UNIVERSIDADE FEDERAL DE SANTA MARIA
CENTRO DE TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Marcos Roberto de Brun Lucca

INTEGRAÇÃO DE TÉCNICAS DE RACIOCÍNIO BASEADO EM CASOS E AGRUPAMENTO DE DADOS NA CONSTRUÇÃO DE SISTEMAS INTELIGENTES

Santa Maria, RS
2017

Marcos Roberto de Brun Lucca

INTEGRAÇÃO DE TÉCNICAS DE RACIOCÍNIO BASEADO EM CASOS E AGRUPAMENTO DE DADOS NA CONSTRUÇÃO DE SISTEMAS INTELIGENTES

Dissertação apresentada ao Curso de Mestrado do Programa de Pós-Graduação em Ciência da Computação (PPGCC), Área de Concentração em Computação, da Universidade Federal de Santa Maria (UFSM, RS), como requisito parcial para obtenção do título de **Mestre em Ciência da Computação**.

Orientador: Profº. Dr. Luís Alvaro de Lima Silva

Santa Maria, RS
2017

Ficha catalográfica elaborada através do Programa de Geração Automática
da Biblioteca Central da UFSM, com os dados fornecidos pelo(a) autor(a).

Lucca, Marcos Roberto de Brun
Integração de Técnicas de Raciocínio Baseado em Casos e Agrupamentos de Dados na Construção de Sistemas Inteligentes / Marcos Roberto de Brun Lucca.- 2017.
110 p.; 30 cm

Orientador: Luís Alvaro de Lima Silva
Dissertação (mestrado) - Universidade Federal de Santa Maria, Centro de Tecnologia, Programa de Pós-Graduação em Ciência da Computação, RS, 2017

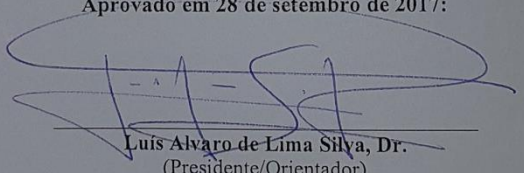
1. Agrupamento de Dados 2. Raciocínio Baseado em Casos 3. Indexação de Casos 4. Indexação de Bases de Casos I. Silva, Luís Alvaro de Lima II. Título.

Marcos Roberto de Brun Lucca

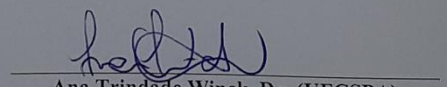
INTEGRAÇÃO DE TÉCNICAS DE RACIOCÍNIO BASEADO EM CASOS E AGRUPAMENTO DE DADOS NA CONSTRUÇÃO DE SISTEMAS INTELIGENTES

Dissertação apresentada ao Curso de Mestrado do Programa de Pós-Graduação em Ciência da Computação (PPGCC), Área de Concentração em Computação, da Universidade Federal de Santa Maria (UFSM, RS), como requisito parcial para obtenção do título de **Mestre em Ciência da Computação**.

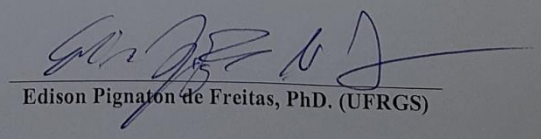
Aprovado em 28 de setembro de 2017:



Luís Alvaro de Lima Silva, Dr.
(Presidente/Orientador)



Ana Trindade Winck, Dr. (UFCSPA)



Edison Pignaton de Freitas, PhD. (UFRGS)

Santa Maria, RS
2017

AGRADECIMENTOS

Agradeço inicialmente ao meu pai Edemar Roberto Lucca, por ter me educado, por ter me dado força, por ter me demonstrado que a vida é uma aventura e, embora não esteja mais presente, agradeço por ser a minha grande inspiração.

A minha mãe Norma Mabel de Brun Lucca, por me educar sempre mostrando que a felicidade está nos atos simples, me ajudando sempre, me mostrando que embora existam obstáculos, sempre há como ser feliz.

A minha irmã Karina Paola de Brun Lucca, por ter sempre estado ao meu lado nas horas mais difíceis e por ser a minha grande amiga.

Ao meu orientador, Prof. Luís Alvaro de Lima Silva, pela oportunidade de executarmos este trabalho juntos, pela transmissão do conhecimento, por me possibilitar crescer como acadêmico e pelas motivações educacionais e pessoais. A oportunidade de concretizar este mestrado é algo que sempre me lembrarei. Obrigado pela amizade.

Ao Prof. Carlos Heitor Moreira pelo tempo dedicado, pela oportunidade de pesquisa em conjunto e pelo apoio para ajudar este trabalho a ser concretizado.

A todos os familiares e amigos que sempre me apoiaram a seguir em minha caminhada.

A UFSM que me acolheu e me deu suporte para o crescimento profissional e acadêmico.

RESUMO

INTEGRAÇÃO DE TÉCNICAS DE RACIOCÍNIO BASEADO EM CASOS E AGRUPAMENTO DE DADOS NA CONSTRUÇÃO DE SISTEMAS INTELIGENTES

AUTOR: Marcos Roberto de Brun Lucca
ORIENTADOR: Luís Alvaro De Lima Silva

Sistemas de Raciocínio Baseado em Casos (Case-Based Reasoning – CBR) utilizam experiências passadas, ou casos, para a tomada de decisão de problemas atuais. Para recuperar essas experiências de bases de casos, técnicas de similaridade são empregadas na busca de casos passados similares ao problema atual usado como consulta. Nestes sistemas, problemas relacionados a análise da relevância de atributos que representam informações gravadas em casos podem impactar na performance de mecanismos de recuperação de casos de CBR. Neste contexto, esta dissertação utiliza algoritmos de clustering aplicados sobre casos para permitir identificar quais são os atributos mais relevantes representados em casos para CBR, os quais podem ser explorados como índices na construção de funções de similaridade ajustadas para a solução de problemas de aplicação. Neste processo, diferentes algoritmos de clustering e métricas de avaliação de grupos de casos são explorados em um framework de indexação, permitindo identificar quais são os atributos mais relevantes. Esta dissertação também utiliza os grupos formados em clustering para a criação de sub-bases de casos a serem utilizadas pelos mecanismos de consultas CBR. Para avaliar a abordagem proposta, um estudo de caso envolvendo um sistema de simulação é explorado. Resultados obtidos neste estudo de caso demonstram que a utilização do framework proposto nesta dissertação permite melhorar a acurácia de consultas CBR realizadas de 44.50% para 83.93%. Bases de casos da Web também são exploradas na validação das propostas apresentadas nesta dissertação.

Palavras-chave: Agrupamento de dados. Raciocínio Baseado em Casos. Indexação de Casos. Indexação de Bases de Casos.

ABSTRACT

INTEGRATION OF CASE-BASED REASONING TECHNIQUES AND CLUSTERING IN THE CONSTRUCTION OF INTELLIGENT SYSTEMS

AUTHOR: Marcos Roberto de Brun Lucca
ADVISOR: Luís Alvaro De Lima Silva

Case-Based Reasoning (CBR) systems use past experiences, or cases, to make decisions about current problems. To retrieve these case base experiments, similarity techniques are employed on the search for past cases similar to the current query problem. In these systems, problems related to the relevance analysis of attributes that represent information recorded in cases can impact the performance of CBR case retrieval mechanisms. In this context, this dissertation uses clustering algorithms applied on cases to identify which are the most relevant attributes represented in CBR case retrieval, which can be exploited as indexes in the construction of similarity functions adjusted for the solution of application problems. In this process, different clustering algorithms and evaluation metrics of clustering cases are investigated in an indexing framework, allowing to identify which are the most relevant attributes. This dissertation also uses clusters formed in clustering to create subcase's bases to be used by CBR query engines. To evaluate the proposed approach, a study case involving a simulation system is explored. Results obtained in this study case demonstrate that the use of the framework proposed in this dissertation allows to improve the accuracy of CBR queries from 44.50% to 83.93%. Web case bases are also explored in the validation of the proposals presented in this dissertation.

Keywords: Clustering. Case-based Reasoning. Case Indexing. Case Base Indexing.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

CBR	Case-Based Reasoning
ESM	The European Simulation and Modelling Conference
IA	Inteligência Artificial
K-NN	k-Nearest Neighbors
MOE	Misture of Experts
PPGCC	Programa de Pós-Graduação em Ciência da Computação
QTD	Queued Top-Down Cluster-Based Retrieval
SIS-ASTROS	Sistema de Simulação Astros 2020
UFSM	Universidade Federal de Santa Maria

SUMÁRIO

1 INTRODUÇÃO

Em Inteligência Artificial (IA), Raciocínio Baseado em Casos (Case-Based Reasoning – CBR) (KOLODNER, 1992) é um paradigma para a solução de problemas baseado na solução de experiências passadas, ou casos. A busca por essas soluções é possível a partir de computações de similaridade entre a descrição do problema atual e as descrições dos casos passados, os quais são tradicionalmente armazenados em bases de casos. Funções de similaridade são geralmente usadas para computar a similaridade entre pares de casos, onde os atributos presentes no corpo dos casos descrevem as características necessárias para o cálculo de similaridade. Em muitas aplicações, alguns atributos utilizados na representação de casos podem ter maior relevância que outros na solução de problemas. Portanto, para que consultas CBR sejam efetivas na recuperação de casos similares de uma base de casos, a utilização de funções de similaridade ajustadas para a solução de problemas de aplicação selecionados é fundamental. Em muitos sentidos, essas funções podem ser ajustadas para refletir a relevância dos atributos utilizados nas computações de similaridade entre casos nestas aplicações.

Para analisar a relevância de atributos utilizados na representação de casos, ou o "peso" destes atributos em computações de similaridade entre casos, esta dissertação investiga a utilização de diferentes técnicas de clustering (JAIN, A.K.; MURTY; FLYNN, 1999). Embora paradigmas que utilizam processos de aprendizado de máquina terem apresentado bons resultados em relação a identificação de atributos mais e menos significativos em computações de similaridade, como os paradigmas que utilizam *introspective learning* (BONZANO; CUNNINGHAM; SMYTH, 1997), *fuzzy indexing* (JENG; LIANG, 1995) e *self-organizing maps and learning vector quantization* (KIM; HAN, 2001) para a indexação de casos, a pesquisa apresentada nesta dissertação explora a utilização de técnicas de clustering pois tais técnicas e seus resultados por si só são de interesse de usuários em muitas aplicações. Entre outros motivos, técnicas de clustering permitem a análise exploratória de dados para descoberta de conhecimento, onde padrões existentes em grandes volumes de dados podem ser revelados, independentemente de qualquer necessidade relacionada à construção de funções de similaridade ajustadas para sistemas CBR. Além disso, existem trabalhos na literatura voltados para a indexação de sistemas CBR via algoritmos de clustering (ARMENGOL, 2011; ARSHADI; JURISICA, 2005; HONG; LIOU, 2008; SILVA, 2010). Nesses trabalhos, a integração de clustering com CBR tem focalizado a definição e utilização de sub-bases de casos geradas por grupos de casos oriundos da execução de algoritmos de clustering, apoiando a construção e utilização de mecanismos de recuperação de casos mais eficientes (MITTAL;

SHARMA; DALAL, 2014; MÜLLER; BERGMANN, 2014; VERNET; GOLOBARDES, 2003; YANG; WU, 2000). Apesar desses trabalhos, na medida do nosso conhecimento algoritmos de clustering ainda não foram diretamente explorados na análise de pesos para atributos e consequente construção de funções de similaridade ajustadas para sistemas CBR, assim como proposto nesta dissertação.

No contexto de CBR, clustering permite agrupar conjuntos de casos oriundos de bases de casos, de tal forma que casos similares sejam organizados no mesmo grupo. Nesta dissertação, a investigação de índices para sistemas CBR é baseada na utilização de clustering na proposição de um novo framework, onde um processo de execução de algoritmos de clustering e avaliação de grupos de casos resultantes destes algoritmos é proposto. Nesse framework, portanto, é executada uma sequência de atividades que envolvem a execução de algoritmos de clustering selecionados: algoritmo baseado em densidade, algoritmo hierárquico e algoritmo particional. A partir disso, é realizada uma avaliação dos grupos gerados por esses algoritmos utilizando as métricas de "entropia", "precisão" e "pureza" (HARTIGAN, 1975), assim permitindo analisar a relevância de atributos usados em computações de similaridade entre casos. Em específico, como algoritmo baseado em densidade o DBScan foi escolhido por possibilitar a visualização de grupos em densidades, como algoritmo hierárquico o DIANA foi escolhido por permitir uma melhor visualização dos grupos gerados via dendrogramas e por ser um algoritmo tradicional de agrupamento hierárquico divisivo e como algoritmo particional o K-Means foi escolhido por ser um dos mais tradicionais algoritmos de clustering utilizados e também por permitir construir uma determinada quantidade de grupos. Em geral, diferentes algoritmos de clustering e diferentes métricas de avaliação de grupos são explorados neste trabalho visando analisar se o processo de análise de índices proposto pode ser baseado em algoritmos de clustering e métricas de avaliação de diferentes naturezas.

1.1 OBJETIVO GERAL

O objetivo geral desta dissertação é explorar algoritmos de clustering no suporte à análise de índices para sistemas CBR visando melhorar a acurácia de consultas realizadas nestes sistemas. Além de explorar a formação e utilização de sub-bases de casos formadas quando algoritmos de clustering são executadas, a ideia é explorar formas de ajustar funções de similaridade de forma que essas funções permitam a melhor resolução de problemas em aplicações sendo tratadas em sistemas CBR. Neste contexto, esta dissertação visa apresentar e discutir atividades pertinentes à proposição de um framework que envolva a execução de

algoritmos de clustering e avaliação de grupos de casos visando a análise de índices a serem explorados em sistemas CBR. Para avaliar as propostas apresentadas neste trabalho, um estudo de caso em um problema de aplicação deve ser realizado. Similares avaliações utilizando bases de casos disponíveis na Web também devem ser realizadas.

1.2 OBJETIVOS ESPECÍFICOS

Objetivos específicos são os seguintes:

a) Explorar algoritmos de clustering para identificar grupos de casos para serem usados como índices de sistemas CBR;
b) Investigar como utilizar resultados das avaliações de qualidade de grupos de casos como "pesos" para atributos na construção de funções de similaridades ajustadas para problemas de aplicação sendo considerados;
c) Implementar algoritmos de clustering de diferentes naturezas; métricas de avaliação de qualidade de grupos de casos juntamente de um processo de transformação dos resultados destas métricas em pesos; organização de sub-bases de casos geradas a partir de grupos formados em clustering para serem usadas em consultas CBR; e funções de avaliação de acurácia de consultas para avaliar a performance de sistemas CBR;
d) Desenvolver um estudo de caso real executando todo o processo de execução de algoritmos de clustering e avaliação de grupos de casos resultantes dessas execuções. O objetivo disso é avaliar se as estruturas de índices criadas permitem melhorar a acurácia de um sistema CBR de aplicação sendo considerado;
e) Executar todo o processo de execução de algoritmos de clustering e avaliação de grupos de casos resultantes dessas execuções utilizando diferentes bases de casos disponíveis na Web para avaliar se as estruturas de índices criadas permitem melhorar a acurácia de diferentes sistemas CBR.

1.3 VISÃO GERAL DAS PROPOSTAS APRESENTADAS NESTA DISSERTAÇÃO

Esta dissertação apresenta um framework para apoiar a análise e utilização de estruturas de índices em sistemas CBR. Este framework é apresentado em um diagrama de atividades na Figura 1. Nesse diagrama, um processo de clustering é apresentado seguido de um processo de

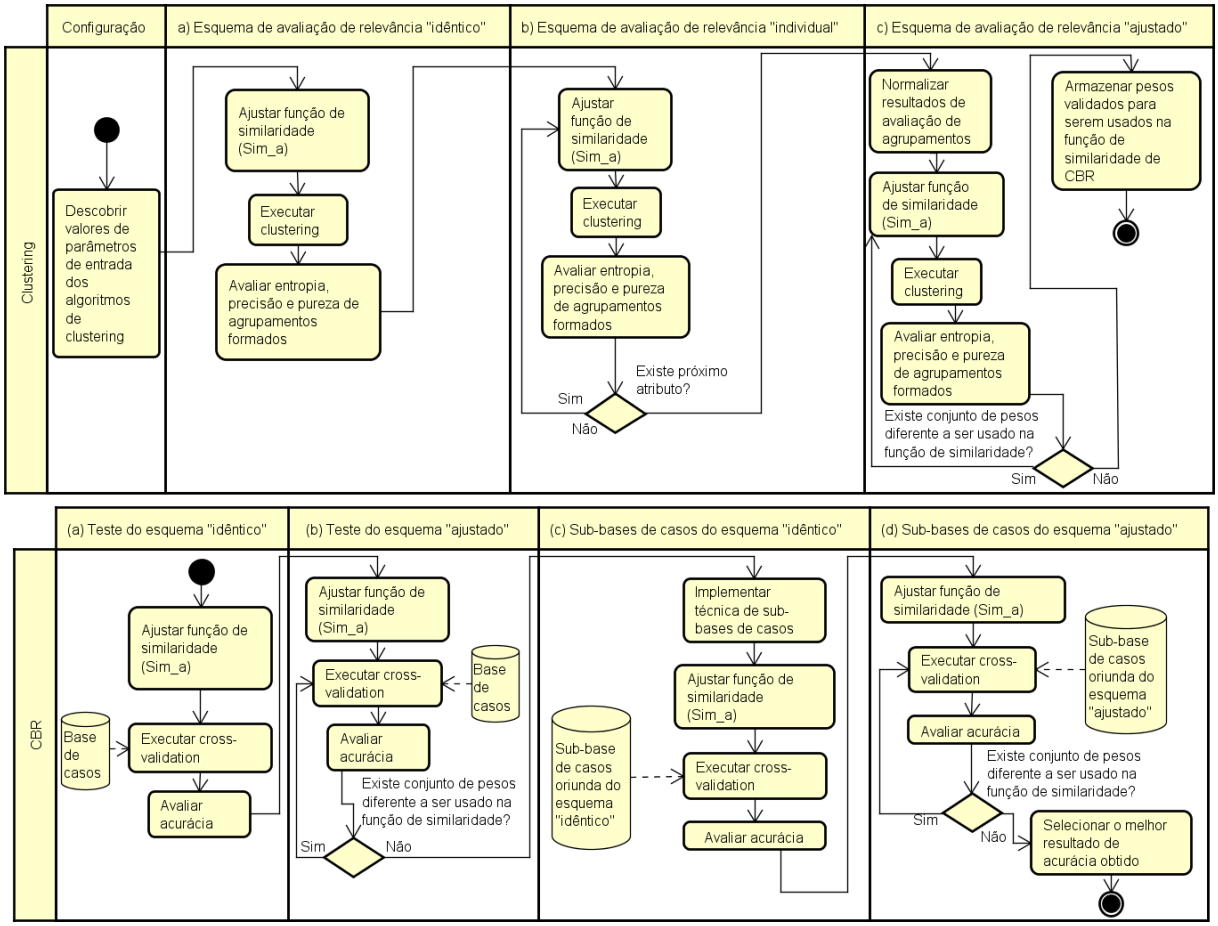
CBR. Em clustering, o objetivo da primeira atividade é executar e avaliar algoritmos de clustering tal como eles são explorados em um processo tradicional de descoberta de conhecimento. A partir dos resultados desses algoritmos, a ideia é identificar valores baseline os quais permitam realizar uma análise qualitativa das estruturas de índice propostas como parte da execução deste framework. A atividade seguinte busca executar e avaliar algoritmos de clustering para obter feedback a respeito de cada um dos atributos usados nas computações de similaridade entre casos. A partir dessa atividade, clustering é explorado de forma não convencional no framework proposto. Neste caso, funções de similaridade utilizadas em clustering são ajustadas onde cada um dos atributos utilizados na representação de casos é individualmente associado a "um valor de peso alto" nas computações de similaridade. Ao usar essa técnica, o impacto desses atributos altamente ponderados na função de similaridade pode ser analisado nos grupos de casos resultantes da utilização dessa função ajustada. O objetivo da última atividade deste framework é explorar o feedback relacionado à relevância de cada um dos atributos de casos na configuração da função de similaridade. A partir disso, é possível executar e avaliar os grupos gerados e comparar os resultados obtidos com resultados baseline.

Em CBR, os índices criados a partir da execução dos algoritmos de clustering e avaliação dos resultados de grupos obtidos são explorados. A primeira atividade deste processo visa avaliar a acurácia de consultas CBR utilizando a função de similaridade não-ajustada, onde todos os atributos são anexados a um "valor de peso igual" nas computações de similaridade. A partir dessa atividade, é possível descobrir um valor baseline de acurácia do sistema CBR utilizado. Em seguida, o processo visa explorar cada conjunto de índices oriundos de clustering como ajuste da função de similaridade. Nas etapas seguintes, o processo de CBR visa explorar as sub-bases de casos obtidas com as execuções dos algoritmos de clustering. Neste caso, a acurácia obtida com as sub-bases de casos oriundas de uma execução tradicional de clustering é analisada, seguida da análise de acurácia obtida utilizando as sub-bases de casos oriundas da execução de clustering com a função de similaridade ajustada.

Para a criação deste framework, implementações foram realizadas para a estruturação dos algoritmos de clustering e das métricas de avaliação de qualidade de grupos formados em clustering. O processo de geração de sub-bases de casos a partir de grupos formados em clustering também foi implementado. Além disso, um sistema CBR foi estruturado para permitir que avaliações de acurácia de consultas por casos similares fossem realizadas. Esta dissertação também apresenta um estudo de caso real onde o framework proposto é executado permitindo obter índices para apoiar a construção do sistema CBR envolvido neste estudo de

caso. Além disso, avaliações do framework são realizadas utilizando bases de casos disponíveis na Web (LICHMAN, 2013).

Figura 1 – Diagrama de atividades do framework de integração entre CBR e clustering.



1.4 ORGANIZAÇÃO DO TEXTO

O texto desta dissertação está organizado em 6 capítulos. O Capítulo 2 apresenta uma revisão bibliográfica sobre o ciclo de vida de CBR, o algoritmo K-NN de CBR, indexação de bases de casos, avaliação de resultados de consultas CBR, o ciclo de vida de clustering, os algoritmos de clustering de densidade, hierárquico e particional, formas de avaliação de resultados de grupos de dados, os trabalhos relacionados à exploração de técnicas de seleção de atributos de casos e, por fim, os trabalhos relacionados à exploração de organização de bases de casos. O Capítulo 3 descreve e documenta um processo de execução de algoritmos de clustering e avaliação de qualidade de grupos de casos resultantes dessas execuções. O Capítulo 4 descreve um estudo de caso realizado para avaliação do framework proposto e também discute

os resultados obtidos. Esse capítulo também demonstra resultados de avaliações utilizando bases de casos disponíveis na Web e também discute estes resultados. O Capítulo 5 apresenta uma discussão sobre a comparação entre os trabalhos relacionados da literatura com o framework desenvolvido nesta dissertação. Por fim, o Capítulo 6 apresenta as conclusões obtidas com a pesquisa desenvolvida e as perspectivas de trabalhos futuros.

2 REVISÃO BIBLIOGRÁFICA

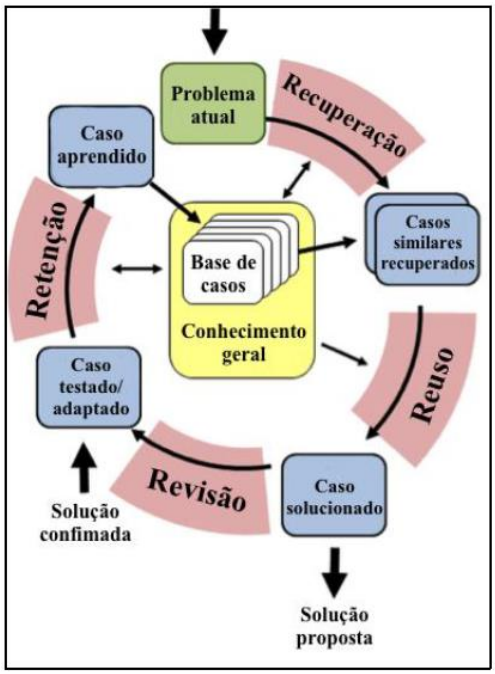
Este capítulo apresenta uma revisão de técnicas de CBR e clustering relevantes para esta dissertação. Este capítulo também revisa trabalhos envolvendo a integração dessas técnicas, assim visando caracterizar as contribuições alcançadas neste trabalho.

2.1 CBR

Os principais repositórios de conhecimento em sistemas CBR são o "vocabulário" utilizado na representação de casos (por exemplo, a identificação de atributos, listas de valores que estes atributos podem conter etc), as medidas de similaridade empregadas na recuperação de casos (por exemplo, medidas ajustadas para melhor avaliar a similaridade entre dois valores de um mesmo atributo entre dois casos etc), o conhecimento que permite a adaptação de soluções passadas para a solução de novos problemas (por exemplo, as regras que definem como uma solução passada pode ser ajustada para a solução de um novo problema) e a base de casos utilizada por um sistema (por exemplo, contendo uma variedade de experiências concretas de solução de problemas) (RICHTER, 1995).

Tradicionalmente, etapas de recuperação de casos, reuso de soluções passadas, revisão da correção de soluções propostas e retenção dessas experiências de solução de problemas na memória caracterizam o desenvolvimento de sistemas CBR (AAMODT; PLAZA, 1994; KOLODNER, 1992). Em linhas gerais, a Figura 1 apresenta uma caracterização desse processo de CBR. Para obter uma resposta para uma consulta em sistemas CBR, uma das etapas mais importantes é a "recuperação de casos". Em muitas aplicações, esta etapa é baseada no emprego de funções de similaridade ajustadas para a solução de um problema sendo investigado. Por este e outros motivos, as etapas de reuso, revisão e retenção de casos não serão exploradas nesta dissertação.

Figura 2 – Ciclo de CBR.



Fonte: AAMODT; PLAZA (1994).

2.1.1 Representação do domínio

A representação do domínio é uma etapa impactante em resultados de consultas CBR. Entre outros problemas, fatos ocorridos no passado costumam ter suas informações dispersas quando armazenados em muitos sistemas computacionais. Embora bancos de dados relacionais sejam amplamente utilizados, a representação de fatos ocorridos no passado deve ser organizada de forma que auxiliem na resolução de problemas correntes, propondo soluções e explicações para consultas realizadas. Dessa forma, bases de casos são utilizadas para capturar fatos ocorridos no contexto de casos passados, os quais são então utilizados por mecanismos de raciocínios implementados em sistemas CBR. Em muitos sistemas CBR, a representação de casos é realizada de forma bastante simplificada, onde estes são capturados por pares (problema, solução). Problemas são especificados por um conjunto de características, capturadas por atributos e valores. Por sua vez, soluções também podem ser representada por um conjunto de atributos e valores.

A escolha de quais atributos devem ser usados na captura de experiências de solução de problemas passados não é simples. Se os atributos escolhidos forem pouco diagnosticados, muitas vezes, eles podem descrever informações superficiais sobre um problema. Neste caso,

tarefas de classificação envolvendo estes casos podem não alcançar resultados satisfatórios. Em geral, a identificação destes atributos busca capturar informações relevantes para um problema para que sistemas CBR consigam responder consultas baseadas na análise de similaridade que é realizada sobre esses conjuntos de atributos e valores representados no corpo de casos.

2.1.2 Cálculo de Similaridade

Com um conjunto de atributos escolhido na etapa de representação de casos, cálculos de similaridade podem ser realizados. Em geral, este cálculo visa computar a similaridade entre casos. Esta similaridade é calculada de acordo com cada atributo presente nos casos, as quais caracterizam as *similaridades locais* associadas a cada um dos atributos representados em casos. Uma vez que similaridades locais tenham sido computadas, *similaridades globais* devem ser também computadas. Essas similaridades globais são medidas no nível de casos, onde tais computações envolvem a utilização de formas de agregação envolvendo resultados de similaridades locais computadas. Normalmente, sistemas CBR utilizam *funções de similaridade ajustadas* para as necessidades de resolução de problemas em aplicações sendo consideradas. Tais ajustes buscam expressar o quanto um atributo pode ser impactante nas computações de similaridade entre casos. Para definir a relevância de cada atributo, normalmente, utiliza-se valores numéricos de pesos.

Em resumo, medir a similaridade entre casos visa comparar uma situação atual com uma situação do passado. Para isso, cada caso existente em uma base de casos pode ser comparado individualmente ao caso atual. Essa comparação é realizada a partir da utilização de uma função de similaridade ajustada para as necessidades de recuperação de casos de um problema de aplicação. Em geral, a definição de quando um caso é similar a outro depende dos atributos utilizados nessas computações de similaridade. Além disso, configurações nestes mecanismos de consulta podem ser realizadas em muitos sistemas CBR. Tais configurações podem estar relacionadas à definição do *número máximo de casos similares a serem recuperados* da base de casos. Também podem estar relacionadas à definição de um *threshold* de recuperação, o qual indica um valor limiar de similaridade mínimo que deve ser alcançado por cada caso recuperado para uma consulta dada. Em situações em que casos recuperados possuem um valor de similaridade em relação a um problema atual que é menor que esse limiar, estes casos não são utilizados no processo de raciocínio implementado por muitos sistemas CBR.

Em sistemas CBR, uma função de similaridade é empregada na computação de similaridade entre pares de casos. Um exemplo clássico deste tipo de função é apresentado na Equação (1):

$$\text{Similaridade}(T, S) = \sum_{k=1}^{p} f(T_k, S_k) \times w_k \quad (1)$$

Nesta equação, T representa o caso atual e S representa o caso passado. f é a função de similaridade utilizada, p é o número de atributos utilizados nestas computações de similaridade e w_k é a estimativa de relevância, ou peso, do atributo k utilizado nestas computações.

Mais uma vez, tais funções de similaridade comumente requerem ajustes para melhor expressar similaridades entre casos. Estes ajustes normalmente se referem a um valor numérico que representa a relevância de cada atributo, ou peso do atributo, utilizado nas computações de similaridade entre casos.

2.1.3 O algoritmo K-NN

Sistemas CBR tipicamente são desenvolvidos utilizando o algoritmo K-NN (*k-Nearest Neighbors*) (FUKUNAGA; NARENDRA, 1975). Este algoritmo possui como parâmetro de entrada a quantidade de casos similares a serem analisados. Neste caso, se o algoritmo K-NN é configurado com k = 1, o caso mais similar ao caso atual é recuperado de uma base de casos. Sendo assim, a solução contida neste caso mais similar recuperado pode ser reutilizada na solução do problema corrente. Quando a quantidade de casos similares é configurada no algoritmo K-NN com um valor maior que 1, como k = 10, 10 casos mais similares podem ser recuperados de uma base de casos. Neste caso, pode-se realizar o voto de maioria dos casos recuperados na busca de uma solução para o problema utilizado como consulta.

2.1.4 Indexação de bases de casos

A seleção e análise da relevância de atributos utilizados na representação de casos, bem como nas computações de similaridade entre casos, além da organização de bases de casos em sub-bases de casos, estão associadas à melhoria de performance e acurácia de sistemas CBR. Neste contexto, problemas normalmente encontrados se referem à decisão do que um caso deve representar. Neste cenário, informações que capturam esses casos devem estar estruturadas de

forma a apoiar a recuperação de casos para a melhor resolução de consultas CBR, assim como descrito em AAMODT; PLAZA (1994).

Um importante aspecto que pode influenciar na perda de acurácia em respostas de consultas CBR é o uso de funções de similaridade pouco expressivas. Neste caso, tais funções podem não representar adequadamente regras existentes em um domínio, não alcançando bons resultados em cálculos de similaridade entre casos. Além disso, quando bases de casos contêm um grande número de casos, consultas realizadas podem não ter um bom desempenho em termos de tempo de computação. Isso deve-se ao custo das computações de similaridade realizadas entre uma consulta e todos os casos armazenados em uma base de casos.

Para atacar esses problemas associados à acurácia e ao desempenho computacional de consultas CBR, métodos de indexação podem ser explorados no desenvolvimento de sistemas CBR. Além de organizar bases de casos em estruturas não planas, onde casos são indexados de várias formas, métodos de indexação objetivam selecionar atributos e analisar a relevância destes de forma a apoiar a criação de funções de similaridade ajustadas para as necessidades de recuperação de casos em aplicações de CBR sendo tratadas. Em geral, tais índices devem ser estruturados para auxiliar nos processos de tomada de decisão. Ao subdividir bases de casos com o uso de índices, portanto, devem ser mantidas as associações existentes entre casos tornando estas sub-bases de casos próprias para consultas. Neste cenário, sub-bases de casos devem conter casos que representem as informações mais comuns entre os seus vizinhos, para que investigações consigam melhor decidir qual a sub-base de caso é a melhor candidata a conter resultados para consultas emitidas em sistemas CBR. Mais ainda, ao indexar atributos utilizados na representação e computação de similaridades entre casos, deve-se manter a representatividade destes para que exista uma fácil compreensão das informações contidas nesses casos (KOLODNER, 1992; LEAKE, 1996).

2.1.5 Avaliação de resultados de CBR

Para avaliar o resultado de consultas em sistemas CBR, medidas de qualidade podem ser utilizadas. Para isto, diferentes medidas podem ser utilizadas, como *precision* e *recall*. A precisão avalia os casos recuperados que são relevantes, enquanto que *recall* avalia os casos relevantes que são recuperados. Nesta dissertação, a qualidade do resultado de consultas CBR é avaliado, assim como definido na Equação (2):

$$\text{Precisão} = \sum \frac{x_i}{x_{ij}} \qquad (2)$$

Nesta equação, x_i são os casos classificados corretamente e x_{ij} é o somatório dos casos classificados corretamente e incorretamente.

O método *leave-one-out cross-validation* também pode ser utilizado para avaliar sistemas CBR. Neste caso, seleciona-se um caso da base de casos por vez e realiza-se uma consulta CBR referente a tal caso, computando a sua acurácia desta consulta. Isso é então realizado até que todos os casos contidos na base de casos tenham sido utilizados como consulta. Sendo assim, o número de testes é igual ao número de casos contidos na base de casos.

2.2 CLUSTERING

Clustering tem como objetivo investigar como organizar dados em grupos utilizando as características naturais destes dados (JAIN, A.K. et al., 1999). A descoberta de grupos existentes entre os dados a partir de clustering permite que usuários realizem diferentes inferências a partir dos grupos obtidos. Ao agrupar objetos similares, consegue-se organizar bases de dados e demonstrar a usuários raciocínios sobre padrões que podem existir a partir dos grupos formados.

No cenário desta dissertação, clustering é utilizado para agrupar casos de bases de casos. Em agrupamentos de casos realizados, um dos objetivos deste trabalho é investigar a relevância dos atributos utilizados na representação de casos e nas computações de similaridade entre eles. Para obter estimativas numéricas relacionadas à relevância de atributos de casos na solução de problemas de aplicação de dados, esta dissertação explora a realização de diferentes configurações em algoritmos de clustering utilizados. A ideia é que essas configurações, principalmente aplicadas a pesos associados a atributos utilizados em computações de similaridade entre casos, consigam demonstrar a influência ou impacto de cada atributo no grupo de casos obtidos quando algoritmos de clustering são executados.

2.2.1 Etapas de clustering

A técnica de clustering pode ser dividida em diferentes etapas. Nesta dissertação, as etapas de seleção de atributos, cálculo de similaridade, agrupamento de casos e avaliação de resultados são exploradas (JAIN, A.K. et al., 1999).

2.2.2 Seleção de atributos

Quando muitos atributos são considerados na representação de casos, um baixo desempenho computacional pode ocorrer na formação de grupos de casos. Devido à falta de índices para esses casos, visto que muitos atributos são utilizados na representação deles, grupos de casos não tão homogêneos podem ser formados. Em muitos sentidos, tais grupos podem não atender às necessidades de recuperação de casos em muitos sistemas CBR. Neste cenário, indexação envolve a seleção de atributos que sejam relevantes para representar casos em bases de casos. Isso não é uma tarefa simples, embora a identificação e utilização destes atributos no processo de agrupamento de casos possa contribuir na formação de grupos homogêneos obtidos a partir da execução de algoritmos de clustering.

Peculiaridades relacionadas a informações capturadas por esses atributos também existem. Por exemplo, atributos podem possuir valores incorretos, como também atributos podem não conter valores. Em um processo de formação de grupos, *outliers* são casos com características dissimilares aos seus vizinhos, sendo distantes dos outros e ocasionando na formação de um grupo próprio para si. Técnicas para a solução destas particularidades podem ser utilizadas, podendo englobar desde modificação, substituição, adição e exclusão de informações de casos.

2.2.3 Cálculo de similaridade

A técnica de clustering depende da similaridade existente entre dados. No contexto dessa dissertação. Funções de similaridade são usadas para calcular a similaridade entre pares de casos. O objetivo do cálculo de similaridade é encontrar um valor numérico que demonstre a relação que um caso possui com o outro. Nas computações de similaridade, os atributos e os tipos dos valores destes atributos devem ser considerados. Funções de similaridade devem ser ajustadas para estas informações. Para a realização da computação de similaridade costuma-se utilizar funções de similaridade tal como a distância Euclidiana, distância de Mahalanobis e distância de Manhattan (MICHALSKI; STEPP, 1983). A distância Euclidiana é tradicionalmente utilizada em muitas aplicações. Ela é caracterizada pela soma da raiz quadrada da diferença entre dois casos, assim sendo baseada no teorema de Pitágoras. Nesta dissertação, a distância Euclidiana ponderada é utilizada, assim como descrita na Equação (3):

$$D(A,B) = \sqrt{\sum_{k=1}^{n} w_k * (A_k - B_k)^2} \qquad (3)$$

Nesta equação, computa-se o quadrado da similaridade entre pares de casos e multiplica-se este resultado pelo peso atribuído aos atributos do caso.

2.2.4 Grupos de casos resultantes de algoritmos de clustering

Grupos de casos possuem características similares entre si, visto que o objetivo de clustering é formar grupos homogêneos entre si. Centroides são objetos que contêm as características comuns a todos os objetos dentro de um grupo. Por exemplo, a cada alocação de um objeto em um grupo, calcula-se a sua similaridade em relação a todos os centroides definidos. O objetivo é escolher o grupo com a menor similaridade, apoiando o processo de formação desses grupos. Dessa forma, casos similares entre si são agrupados de forma que se tornem vizinhos.

Para cada método de clustering, existem diferentes algoritmos baseados em diferentes regras de conectividade entre os dados para a formação de grupos. Na formação desses grupos, existem algoritmos que usam regras de conectividade, algoritmos que usam centroides, algoritmos que usam modelos de distribuição e algoritmos que usam modelos de densidade. Algoritmos de clustering explorados nesta dissertação são detalhados na Seção 2.2.6.

2.2.5 Avaliação de grupos de casos resultantes da execução de algoritmos de clustering

A avaliação da qualidade de grupos de casos formados normalmente é dependente do domínio de aplicação em que o algoritmo de clustering foi aplicado. Para essa avaliação ocorrer, comumente, é importante a participação de especialistas do domínio de aplicação sendo tratado. Com ou sem auxílio de especialistas de domínio, a avaliação de qualidade de grupos de casos pode ser realizada de três maneiras: avaliação interna, avaliação externa e avaliação relativa (JAIN, A.K. et al., 1999).

No contexto de casos disponíveis em bases de casos, a avaliação externa de grupos de casos toma como referência as soluções representadas no corpo dos casos, permitindo determinar se casos são corretamente classificados ou não. A avaliação interna busca avaliar a homogeneidade dos grupos formados, avaliando a variância que pode existir entre os atributos dos casos. A avaliação relativa compara resultados obtidos com estruturas pré-definidas, avaliando se grupos contêm casos classificados corretamente conforme uma estrutura de grupo pré-definida.

Nesta dissertação, métricas de entropia, pureza e precisão foram exploradas na avaliação externa de grupos de casos obtidos a partir da execução de algoritmos de clustering. É importante notar que tais métricas possuem propósitos diferentes: a pureza objetiva verificar a máxima ocorrência de classes (soluções) nos grupos; a precisão objetiva avaliar se casos recuperados estão corretos de acordo com a solução presente no corpo desses casos, definindo a proporção de casos recuperados corretamente em cada grupo e a entropia objetiva avaliar a desordem entre valores dos atributos dos casos.

2.2.5.1 Precisão

A precisão é uma medida que avalia a quantidade de casos corretamente classificados conforme uma classe específica dentro de um conjunto de casos pertencentes a um grupo de casos. A Equação (4) detalha como a métrica de Precisão pode ser medida:

$$\text{Precisão (p)} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{4}$$

Nesta equação, TP é a representação dos *true-positives* que são os casos corretamente recuperados para consultas dadas, onde um caso utilizado como consulta e um caso mais similar recuperado possuindo a mesma solução do caso utilizado como consulta são contabilizados como *true-positives*. FP é a representação dos *false-positives* que são os casos incorretamente recuperados, assim como avaliados pela comparação das soluções representadas no corpo desses casos. Na equação de precisão (4), deve-se dividir os casos corretamente recuperados pela soma de todos os casos recuperados. Nesta dissertação, medidas de precisão são calculadas para cada um dos grupos de casos formados a partir da execução de algoritmos de clustering. Em seguida, é realizada a média aritmética dessas estimativas de precisões.

2.2.5.2 Entropia

A entropia avalia a diversidade entre casos vizinhos em um mesmo grupo de casos. Esta métrica focaliza a análise da desordem que pode existir neste grupo. A entropia avalia separadamente cada atributo dos casos em relação a seus vizinhos, descobrindo a variância que pode existir em cada grupo. Equação (5) detalha como a métrica de Entropia pode ser medida:

$$\text{Entropia(X)} = -\sum_{i=1}^{n} \text{P}(c_i) * \log_2\big(\text{P}(c_i)\big) \tag{5}$$

Nesta equação, P representa a probabilidade de ocorrência da solução de um caso e $\log_2$ suaviza essa probabilidade de cada atributo da solução do caso ocorrer. Nesta dissertação, a entropia é calculada para cada um dos grupos de casos formados em clustering. Em seguida, a média dessas entropias é calculada. Quanto menor a entropia, maiores os níveis de organização interna dos grupos.

2.2.5.3 Pureza

A pureza verifica a coerência dos grupos de casos formados em relação à máxima ocorrência de casos pertencentes a uma determinada classe. Nesta avaliação, cada grupo de casos é representado pela classe de casos mais frequente naquele grupo. A pureza é calculada dividindo a soma dos casos corretamente atribuídos a uma determinada classe em um grupo pelo número total de grupos. Quanto maior a pureza, mais coesivo e menos acoplado é o grupo de casos. Equação (6) detalha como a métrica de Pureza pode ser medida:

$$\text{Pureza} = \frac{\sum_{i=1}^{n} \frac{|c_i^j|}{|c_i|}}{n} \qquad (6)$$

Nesta equação, $|c_i^j|$ representa a quantidade de casos com a solução calculada, denotando a classe dominante em um grupo de casos i. $|c_i|$ representa o número de casos no grupo de casos i, e n o número total de grupos de casos. Nesta dissertação, a pureza é calculada para cada um dos grupos de casos formados em clustering. Em seguida, a média desses valores de pureza é calculada.

2.2.6 Algoritmos de clustering

Diferentes paradigmas de clustering podem ser encontrados na literatura (HARTIGAN, 1975). Entre outras organizações propostas, técnicas de clustering podem ser organizadas quanto à organização espacial de grupos, como agrupamentos planos e agrupamentos hierárquicos. Em métodos de clustering planos, os grupos podem estar em um mesmo nível, mas separados por classes. Em métodos de clustering hierárquicos, os grupos são organizados em diferentes níveis de taxonomias, sendo necessário definir o nível de corte da hierarquia para a formação de uma determinada quantidade de grupos. Neste caso, produzindo uma maior quantidade de grupos obtém-se uma granularidade mais fina nos grupos, o que pode representar uma maior homogeneidade relacionada aos grupos formados.

Embora algoritmos possam ser classificados de diversas maneiras, devido a estruturas lógicas existentes entre eles, uma forma de categorizar de modo geral os algoritmos é os conceituando como paramétricos e não-paramétricos (FUNG, 2001). O objetivo de utilizar essa definição é porque algoritmos podem utilizar parâmetros como configuração para a formação dos grupos.

A estruturação dos grupos depende diretamente do tipo de algoritmo de clustering escolhido, visto que conforme a escolha, diferentes paradigmas de clustering podem ser considerados. Exemplos de categorias de algoritmos que são considerados nesta dissertação são o modelo de conectividade, o modelo de centroide e o modelo de densidade.

2.2.6.1 Modelos de conectividade

Algoritmos que usam regras de conectividade para a formação de grupos são fundamentados na noção de que pontos de dados mais próximos em um espaço Euclidiano possuem maior semelhança uns com os outros do que os pontos mais distantes. Esses modelos podem ter duas abordagens. Na primeira abordagem, conhecida como aglomerativa, o grupo de casos inicia-se em grupos separados para que então eles sejam associados a grupos homogêneos, possuindo objetos similares à medida que a similaridade entre estes objetos diminui. Na segunda abordagem, conhecida como divisiva, todos os casos são agrupados em um único grupo para que, depois, sejam divididos conforme a similaridade entre os objetos aumente. Algoritmos que usam regras de conectividade são fáceis de serem interpretados, embora não possuam escalabilidade para agrupar grandes volumes de dados. Exemplos destes modelos são os algoritmos AGNES (aglomerativo) e DIANA (divisivo) (KAUFMAN; ROUSSEEUW, 2008).

Em algoritmos hierárquicos, ocorre uma partição sucessiva dos casos contidos em uma base de casos, até que esses casos sejam hierarquicamente representados, por meio de uma árvore (EVERITT, 1974). Em geral, não é necessário definir um número de grupos de casos quando esses algoritmos são executados. Neste caso, é possível definir um ponto de corte, assim determinando um número de grupos, quando uma hierarquia é formada. O método hierárquico é executado com a utilização de uma matriz de similaridades entre os grupos, as quais representam as medidas de similaridade entre os objetos sendo agrupados.

Algoritmos hierárquicos visam à investigação de diferentes níveis de hierarquia associados aos grupos formados, permitindo que usuários obtenham maior clareza na visualização destes. Os algoritmos hierárquicos, em sua maioria, possuem a sua execução

baseada em recursividade. Nestas, prevê-se como entrada grupos identificados brevemente e define-se como saída uma hierarquia de casos agrupados conforme suas medidas de similaridade, dentro de categorias formadas pelo algoritmo. Os parâmetros a serem entrados neste algoritmo são: a) o critério de similaridade para a formação de grupos e b) a quantidade máxima de hierarquias a serem formadas. Os critérios de similaridade objetivam calcular a similaridade entre elementos dentro dos grupos. As medidas de similaridade que normalmente são utilizadas em algoritmos hierárquicos são: single-link, complete-link, average link e outras:

a) Single-link representa a similaridade entre pares de casos mais próximos;
b) Complete-link representa a similaridade entre pares de casos mais distantes;
c) Average link representa a similaridade média entre casos dentro de cada grupo de casos.

A representatividade de grupos utilizando hierarquias, independente da forma que esta hierarquia esteja implementada, define o conceito base de algoritmos hierárquicos. Estes algoritmos objetivam representar as conexões que possam existir entre grupos de casos em um formato de árvore, sendo esta a estrutura de dados que mais se aproxima de uma representação taxonômica de informações.

2.2.6.2 Modelos de centroide

Os modelos de centroide representam algoritmos de agrupamento iterativos em que a noção de similaridade é derivada pela proximidade em que um caso possui em relação ao centroide de seu grupo. O K-Means é um dos algoritmos mais tradicionais desta categoria. Este algoritmo possui como parâmetro básico de entrada a quantidade de grupos a serem formados.

Algoritmos particionais são exemplos de algoritmos pertencentes ao modelo de centroide. Conforme FUNG (2001), o método particional é utilizado para minimizar uma função de custo de clustering. Uma questão chave de métodos particionais é a decisão do número de grupos a serem formados. De acordo com KAINULAINEN (2002), uma definição errônea da quantidade de grupos pode implicar em resultados insatisfatórios na classificação de objetos.

O algoritmo K-Means é um dos algoritmos mais simples e utilizados na literatura (JAIN, A.K. et al., 1999). Neste algoritmo, calculam-se valores médios entre todos os casos para formar centroides, que são redefinidos a cada execução do algoritmo. Cada caso que está sendo agrupado deve ter a sua similaridade em relação ao centroide calculada, proporcionando a criação de grupos de casos baseada na menor distância entre elementos. A cada alocação de um

caso em um determinado grupo, todos os centroides devem ser recalculados, para que as alocações consigam colocar estes elementos nos grupos mais similares. A cada iteração deste algoritmo, ele busca uma certa quantidade de grupos a serem formados, podendo submeter objetos a grupos que não são tão similares, pelo simples fato de ter que criar uma quantidade determinada de grupos (definida como parâmetro de entrada). Assim, podem ser formados grupos menos homogêneos ou com uma alta ocorrência de outliers.

2.2.6.3 Modelos de densidade

Algoritmos de clustering baseados em densidade permitem a formação de formas arbitrárias de grupos, devido à conectividade com que elementos podem possuir quando são representados através de densidades (ESTER et al., 1996). Nesse modelo, a densidade representa a similaridade entre casos. Em geral, espaços com maiores aglomerações de dados são mais densos que espaços com menores aglomerações.

Segundo MADHULATHA (2012), o algoritmo DBScan representa densidades através de conjuntos de pontos em um espaço cartesiano, onde tais pontos podem representar casos. Quanto mais próximos estão estes casos, mais forte é a representação de um grupo de casos. Quanto mais fracas as densidades, mais forte é a representatividade de outliers. O algoritmo DBScan possui os seguintes parâmetros básicos de entrada: a) o raio (eps) que define um perímetro a ser agrupado, definindo que somente casos dentro do mesmo raio pertencem a tal grupo de casos; e b) a quantidade mínima de casos para formar os grupos (minpts), condicionando à formação de grupos de casos a uma quantidade de vizinhos superior a este valor previamente estabelecido. Utilizando a configuração destes parâmetros, as coordenadas desses casos são avaliadas em um espaço cartesiano. Este algoritmo costuma ser mais representativo em problemas onde um número inicial de grupos não é estipulado.

O DBScan possui a sua execução inicializada com a seleção de esferas relacionadas a cada caso, conforme o raio, informado como parâmetro de entrada, contando a quantidade de casos que cada esfera contém. Se o número de casos dentro de cada esfera for superior a minpts, marca-se o centro da esfera como pertencente a um grupo. Em seguida, marca-se os casos dentro de tal esfera como pertencentes a mesma aglomeração. Para as outras esferas, recursivamente realiza-se o mesmo processo de formação de grupos. Se a quantidade de casos dentro uma esfera for inferior a minpts, ignora-se a esfera e continua o processo para a esfera seguinte. Outliers podem ser formados no fim do processamento, representando os casos que não geraram grupos. Estes outliers podem estar contidos em outros raios, podendo ser parte de

grupos conforme o algoritmo é executado. O processo de verificação de raio e agregação de casos aos grupos continua até que as aglomerações sejam completamente formadas.

2.3 REVISÃO DE TÉCNICAS DE INTEGRAÇÃO ENTRE CBR E CLUSTERING

Em CBR, soluções para problemas atuais são propostas a partir do reuso de soluções de problemas ocorridos no passado. Para isso, o sucesso de sistemas CBR depende efetivamente da fase de recuperação de casos de bases de casos, a qual envolve a computação de similaridades entre casos. Tarefas de indexação em sistemas CBR buscam apresentar soluções para diferentes problemas relacionados a recuperação de casos e organização de casos em memória. Entre eles, indexação pode envolver identificar a relevância de atributos utilizados em computações de similaridade e estruturar bases de casos visando apresentar respostas para consultas CBR.

Em um trabalho de revisão da literatura a respeito da indexação de sistemas CBR, assim como descrito em WETTSCHERECK; AHA; MOHRI (1997), métodos de indexação para algoritmos de CBR (os quais são algoritmos classificados como *lazy-learning*) (ARMENGOL, 2011) são analisados de acordo com cinco dimensões. A primeira dimensão diz respeito à utilização de técnicas que usam *feedback de usuários* para ajustes de funções de similaridade. A segunda dimensão focaliza o *desempenho de classificações de bases de casos* no que diz respeito ao *tamanho dessas bases*. De forma simples, quando um baixo desempenho é observado, casos irrelevantes podem ser eliminados das bases de casos e, consequentemente, das computações de similaridade. A terceira dimensão diz respeito *à organização do conjunto de atributos* para que, depois de organizados, os atributos possam ser ponderados em funções de similaridade. Nesta organização, valores de atributos errados ou redundantes podem ser removidos dos casos, e os valores de atributos dos casos vizinhos podem ser utilizados. A quarta dimensão analisa a *utilização e impacto de funções de similaridade locais e globais em computações de similaridade*. Neste caso, o trabalho compara o uso de funções de similaridade locais (no nível de atributos tomados individualmente) com funções globais (no nível de casos), constatando que, ao usar funções locais de similaridade, é possível formar grupos de casos mais homogêneos. Isto ocorre quando uma aplicação testada contém regras que, ao serem implementadas através de funções de similaridade locais, apoiam a obtenção de melhores resultados de acurácia decorrentes de computações de similaridade. A quinta e última dimensão analisa *conhecimento relacionado à participação de especialistas na configuração de funções de similaridade*.

Propostas voltadas para a indexação de bases de casos, onde técnicas de clustering são exploradas, são apresentadas na literatura. Tais propostas podem ser divididas em dois tópicos: seleção de atributos de casos a serem utilizados em computações de similaridade e grupo de casos para organização e utilização de sub-bases de casos em sistemas CBR.

2.3.1 Seleção de atributos de casos a serem utilizados em computações de similaridade

No trabalho de HONG; LIOU (2008), a técnica de clustering é usada para coletar explicações a respeito de classes de casos, ou grupos de casos, que podem ser definidas em um problema de aplicação. A ideia é produzir grupos significativos para usuários de aplicações sendo consideradas. Neste contexto, um grande número de atributos pode ser usado na representação de casos, transformando a recuperação de casos em uma tarefa computacional complexa e muitas vezes pouco precisa. Nessas situações, a seleção e consequente indexação de um subconjunto mais relevantes de atributos pode ser necessária em muitos problemas de aplicação.

Em clustering, casos contidos em um mesmo grupo de casos possuem alta similaridade, enquanto casos contidos em diferentes grupos possuem alta dissimilaridade. Em um mesmo grupo de casos, atributos de caso que possuem informações possivelmente anômalas (possivelmente incorretas, por vários motivos) podem ser descartados e substituídos pelas características mais similares provenientes de casos vizinhos. Neste cenário, a técnica de indexação de atributos é adequada às seguintes três situações. Na primeira, todos os valores dos atributos representativos existem na base de casos. Na segunda, alguns valores dos atributos são desconhecidos. Na terceira, alguns atributos representativos não aparecem na base de casos. Na segunda e terceira situações, um processo de substituição de atributos e valores pode ser aplicado para que seja possível computar a similaridade entre pares de casos. Quando é necessário substituir atributos e valores de um caso, o atributo mais similar é utilizado proporcionando às novas informações. Portanto, o enfoque proposto no trabalho de HONG; LIOU (2008) pode evitar que computações de similaridade sejam realizadas em informações possivelmente incorretas, devido à organização de casos em grupos de casos homogêneos. Dessa forma, é possível melhorar a performance de sistemas CBR, onde informações anômalas não são consideradas em computações de similaridade.

No trabalho de ARSHADI; JURISICA (2005), a integração entre CBR e clustering é voltada para a seleção de características a serem utilizadas na representação de casos, de forma a melhorar a acurácia de consultas CBR. Devido à existência da diversidade de dados que

podem existir em aplicações na área da biologia, visto que esta é a aplicação alvo da pesquisa apresentada, a técnica de clustering é utilizada na descoberta de inferências neste domínio. A seleção de características objetiva identificar atributos "informativos", os quais têm relevância na compreensão e identificação de genes. Por outro lado, a remoção de atributos "não-informativos" apoia a obtenção de melhores resultados de acurácia na classificação de casos. Em geral, a abordagem utilizada é baseada na mistura de especialistas (*mixture-of-experts*) para CBR (MOE4CBR). Essa abordagem utiliza a ideia de três componentes integrados: a) um sistema CBR que prediz a classe que um novo caso está relacionado; b) clustering para criação de k grupos de casos e c) técnicas de seleção de características para representar cada grupo a partir da utilização de um conjunto de atributos.

Em CBR, a mistura de especialistas (MOE) é utilizada para predizer as classes de novos casos tomados como consultas. Para isso, um sistema de votação é utilizado, onde os votos são fornecidos por cada especialista. Assim como descrito em ARSHADI; JURISICA (2005), o desempenho desses especialistas pode ser melhorado utilizando técnicas de clustering e técnicas de seleção de características. Para cada especialista, uma sub-base de casos é organizada a partir da execução de clustering, onde cada um dos especialistas possui a tarefa de realizar a classificação de casos de acordo com suas próprias sub-bases de casos. Neste trabalho, os algoritmos de clustering utilizados são K-Means, mapas auto organizáveis e spectral clustering. De acordo com análises realizadas, spectral clustering possui melhor desempenho quando comparado às outras duas técnicas.

Em geral, os trabalhos relacionados aqui revisados não exploram a ideia de melhorar a acurácia de consultas CBR utilizando estimativas de relevância para os atributos de casos, assim como explorado nesta dissertação. Contudo, a ponderação de atributos não deixa de ser uma forma de seleção de atributos usados na representação de casos e computação de similaridades entre casos, assim como descrito nos trabalhos relacionados. Além disso, esta dissertação propõe a construção e análise de grupos de casos criados a partir de diferentes algoritmos de clustering e diferentes métricas de avaliação desses grupos. A ideia disso é explorar o processo de análise de índices proposto a partir do uso de técnicas de diferentes naturezas. Ainda mais, esta dissertação demonstra como descobrir valores baseline de avaliação de resultados dos grupos de casos para usar esses baselines na comparação com outros resultados e gerar pesos para atributos de casos. Assim como apresentados em trabalhos relacionados, técnicas de clustering também são exploradas nesta dissertação na organização e utilização de sub-bases de casos na construção de respostas mais precisas para consultas CBR.

2.3.2 Agrupamento de casos visando a construção de sub-bases de casos

No desenvolvimento de sistemas CBR, diferentes trabalhos demonstram abordagens de indexação onde técnicas de clustering são utilizadas. A abordagem mais comum envolve o uso destas técnicas na análise de como grandes bases de casos poderiam ser divididas em sub-bases de casos. Esta é uma questão relacionada à organização dos casos em memória, assim como investigado em MITTAL et al. (2014), MÜLLER; BERGMANN (2014), VERNET; GOLOBARDES (2003) e YANG; WU (2000). Com sub-bases de casos contendo casos com características semelhantes, a ideia geral é melhorar o desempenho de consultas CBR, uma vez que cálculos de similaridade envolvendo grandes bases de casos podem ser uma tarefa demorada. Na maioria desses trabalhos, as características comuns dos casos são identificadas e representadas por centroides. Conforme identificado quando algoritmos de clustering são usados, esses centroides são definidos como índices dos grupos de casos formados. Esses grupos, então, são tomados como bases de casos distintas nestes sistemas CBR. Dessa forma, consultas CBR priorizam o uso desses índices (ou centroides) na recuperação de casos similares contidos nestas sub-bases de casos, assim como explorado nesta dissertação.

No trabalho de YANG; WU (2000), uma grande base de casos é dividida em grupos de casos menores. O objetivo é manter a estrutura da base de casos simples e a política de manutenção dela transparente, para conseguir uma melhor acurácia e um melhor desempenho computacional na recuperação de casos em CBR. A técnica é baseada em dois métodos. No primeiro, o algoritmo CBSCAN de clustering que é baseado no algoritmo DBScan (YANG; WU, 2000) é usado para dividir a base de casos em sub-bases de casos. No segundo, o usuário participa do processo de agrupamento selecionando atributos relevantes. As sub-bases de casos são distribuídas em diferentes locais de uma rede, onde cada uma destas sub-bases de casos é representada por um centroide capturando as características mais comuns dos casos. Em cada consulta realizada no sistema, os atributos dos centroides são apresentados ao usuário em uma forma interativa, de forma que o usuário pode escolher os atributos de seu interesse. A cada iteração, as sub-bases de casos que não são de interesse do usuário são removidas do processo de consulta. As demais sub-bases de casos então podem ser sugeridas para os usuários conforme a escolha interativa das características dos casos. Esse processo interativo de seleção é repetido até que uma sub-base de casos seja escolhida. Neste ponto, o sistema CBR utiliza a sub-base de casos escolhida na computação de respostas para consultas. No contexto de organização de bases de casos, onde bases de casos podem estar estruturadas de forma complexa dificultando computações de similaridade, o trabalho apresentado em YANG; WU (2000) explora a ideia de

criação de sub-bases de casos a partir de grupos de casos formados em clustering e investigação da participação do usuário na seleção de atributos relevantes dessas sub-bases de casos.

No trabalho de VERNET; GOLOBARDES (2003), os algoritmos Mean Sphere e Mean K-Means são utilizados para apoiar a organização de casos de bases de casos em memória. O objetivo é melhorar o desempenho computacional de consultas CBR e também reduzir o número de outliers dos grupos de casos formados. O enfoque deste trabalho é dividido em dois níveis. No primeiro, esferas (grupos) são construídas de acordo com a distribuição dos casos em memória. No segundo, clustering é executado novamente em cada uma dessas esferas. Neste enfoque, cada caso da base de casos é distribuído em uma esfera, conforme a sua classe associada. Então, centroides são formados para cada esfera contendo valores comuns das características presentes nos casos de cada esfera. O objetivo da formação destes centroides é utilizar eles como índices que apoiem comparações mais rápidas entre um caso usado como consulta e um centroide. Ao executar novamente clustering, as esferas originais são reorganizadas com a aplicação do algoritmo K-Means, o qual é aplicado internamente em cada esfera. Com esta reorganização dos casos em memória, o objetivo é reduzir o tempo computacional da recuperação de casos do sistema CBR. Neste cenário, ao executar uma consulta, o grupo que possui o centroide mais similar ao caso consultado é selecionado. No contexto de indexação de sistemas CBR onde um grande volume de casos está disponível para consulta, esse artigo explora a ideia de utilizar resultados de diferentes algoritmos de clustering de forma combinada na construção de respostas para consultas CBR.

No trabalho de MÜLLER; BERGMANN (2014), a indexação de casos de uma base de casos representando descrições de “processos” é investigada. Nestes sistemas, casos costumam estar representados por estruturas de grafos onde a computação de similaridade costuma ser realizada a partir de comparações entre subgrafos, o que normalmente é computacionalmente custoso. Portanto, melhorar o desempenho da fase de recuperação de casos é uma necessidade nestes sistemas. Nesse trabalho existe então uma preocupação com a eficiência da função de similaridade e com a estrutura de representação dos casos. Na técnica proposta, o algoritmo hierárquico de clustering é usado para a organização dos casos em uma estrutura de árvore. A partir disso, os casos são alocados em grupos de casos, possibilitando a recuperação de casos similares através de uma pesquisa na árvore, visto que os grupos são representados por centroides. O algoritmo hierárquico de recuperação de casos explorado é baseado em uma estrutura top-down, denominada *Queued Top-Down Cluster-Based Retrieval* (QTD). O enfoque é baseado em dois parâmetros, onde o primeiro define um nível superior limite da árvore que organiza a memória de casos, enquanto o segundo define um nível inferior limite da

árvore. Com isso, as consultas são realizadas em grupos de casos que estejam entre estes dois níveis, utilizando um algoritmo de busca heurística. Para isso, um primeiro grupo de casos com a maior similaridade à consulta dada é selecionado. Se esse grupo de casos não for uma folha, os seus grupos filhos (contidos nos nós esquerdo e direito da estrutura hierárquica que representa a base de casos) são investigados. No contexto de indexação de sistemas CBR onde um grande volume de casos relacionado a "processos" está disponível para consulta, o trabalho apresentando em MÜLLER; BERGMANN (2014) explora a ideia de agrupar casos utilizando um algoritmo hierárquico e utilizar esses grupos na computação de respostas mais rápidas para consultas CBR.

Em geral, os trabalhos relacionados aqui revisados não discutem a utilização de funções de similaridade ajustadas para a criação de sub-bases de casos para apoiar mecanismos de consulta CBR, assim como explorado nesta dissertação. Além disso, esta dissertação propõe construir e analisar sub-bases de casos a partir de diferentes algoritmos de clustering e com diferentes ajustes na função de similaridade, onde centroides dos grupos gerados por esses algoritmos de clustering são usados como índices na construção dessas sub-bases de casos. Ainda mais, esta dissertação discute resultados de acurácia obtidos com a execução do mecanismo de consultas de casos similares ajustado segundo as diferentes técnicas de indexação propostas, onde, inclusive, resultados de acurácia obtidos utilizando sub-bases de casos formadas sem utilizar a função de similaridade criada são comparados a resultados de acurácia obtidos usando uma função de similaridade criada.

2.4 CONSIDERAÇÕES DO CAPÍTULO

Em sistemas CBR, a recuperação de casos é uma das etapas mais críticas, pois problemas na organização da base de casos, tal como nas informações contidas nos casos, afetam os resultados de consultas CBR, impedindo que estas consultas sejam respondidas com acurácia e também afetam a compreensibilidade das soluções respondidas a um dado problema. Por isso, é importante organizar a base de casos para identificar, analisar e entender as reais necessidades do domínio. Para que o mecanismo de consulta de casos seja efetivo, ele pode utilizar estruturas de índices de bases de casos, utilizando técnicas externas, como clustering, que também auxilia usuários a descobrir padrões existentes nos dados.

Diferentes técnicas de clustering exploradas nos trabalhos relacionados podem ser aplicadas para desenvolver estas atividades. Dentre as técnicas exploradas, temos a organização de informações contidas no corpo dos casos. Grupos de casos agrupam casos com informações

semelhantes, gerando grupos homogêneos e heterogêneos entre si. Assim, os grupos gerados em clustering podem ser utilizados na criação de índices para servirem como base na correção de informações anômalas dos casos. Temos também a organização de bases de casos em sub-bases de casos como técnica explorada nos trabalhos relacionados. Grupos gerados em clustering referentes a bases de casos organizam casos em estruturas menores, as quais podem ser utilizadas como sub-bases de casos. Além disso, centroides contendo as informações mais comuns dos casos agrupados podem ser utilizados como índices na construção dessas sub-bases de casos. A ideia de utilizar sub-bases de casos é proporcionar a mecanismos de consultas de casos, um meio mais rápido na busca por casos similares, onde a sub-base de caso que contém o índice mais similar ao caso consultado é a utilizada para a realização de tal consulta CBR, assim melhorando o desempenho da etapa de recuperação de casos.

3 INTEGRAÇÃO ENTRE CLUSTERING E CBR

Este capítulo detalha uma sequência de atividades para auxiliar usuários a estimar o peso que cada atributo utilizado na representação de casos deve ter no cálculo de similaridade entre casos. Na investigação desses índices utilizados em CBR, diferentes configurações de pesos são realizadas na função de similaridade utilizada por algoritmos de clustering. A ideia é utilizar essas funções de similaridade ajustadas para as necessidades de recuperação de casos de um problema de aplicação sendo considerado, buscando obter grupos de casos que apoiem a resolução de um problema. Em geral, tais grupos visam refletir diferentes subclasses de problemas, as quais são identificadas a partir da análise dos casos disponíveis em bases de casos para CBR. Além disso, os grupos formados nesse processo de clustering podem apoiar a organização da base de casos em sub-bases de casos para que consultas CBR sejam realizadas nessas sub-bases de casos, assim como descrito em trabalhos relacionados (MITTAL et al., 2014; MÜLLER; BERGMANN, 2014; VERNET; GOLOBARDES, 2003; YANG; WU, 2000).

3.1 PROCESSO DE INTEGRAÇÃO ENTRE CLUSTERING E CBR

A integração entre clustering e CBR pode acontecer de diferentes formas visando à solução de diferentes problemas de indexação em sistemas CBR. Neste contexto, trabalhos apresentados na literatura podem ser identificados (ARMENGOL, 2011; ARSHADI; JURISICA, 2005; HONG; LIOU, 2008; MITTAL et al., 2014; MÜLLER; BERGMANN, 2014; VERNET; GOLOBARDES, 2003; WETTSCHERECK et al., 1997; YANG; WU, 2000). No nosso projeto, essa integração visa a construção de funções de similaridade ajustadas utilizadas como índices no apoio a recuperação de casos em sistemas CBR.

Geralmente, a construção de funções de similaridade para resolver problemas de aplicação de consultas CBR requer a utilização de valores de pesos (ver distância Euclidiana ponderada na Seção 2.2.3), os quais são associados a cada atributo utilizado em computações de similaridade entre casos. Em muitos sistemas, a definição desses pesos requer a participação de usuários com conhecimento no domínio de aplicação sendo considerado. Baseado ou não em especialistas de domínio que orientem esse processo de análise da significância de atributos na solução de problemas dados, a utilização de processos de desenvolvimento (detalhados como sequencias de atividades, assim como explorado em metodologias de desenvolvimento de software) que apoiem a indexação de sistemas CBR é um tópico de pesquisa onde diferentes

propostas são apresentadas na literatura, assim como descrito na Seção 2.3. Diferente dessas propostas, algoritmos de clustering e métricas de avaliação de grupos de casos resultantes desses algoritmos podem ser explorados sistematicamente neste processo de indexação de casos.

Algoritmos de clustering são geralmente utilizados na análise exploratória de dados. Entre outros objetivos, esses algoritmos buscam apoiar a descoberta de padrões que possam existir nestes dados, os quais são revelados por meio de grupos de dados resultantes da execução desses algoritmos. No cenário deste trabalho, ao invés de focalizar a análise de dados, clustering é utilizado para analisar grupos de casos disponíveis em bases de casos. É importante ressaltar que isso é diferente de agrupar dados, os quais são comumente encontrados em repositórios de dados. Entre outros motivos, um caso captura mais informação que um conjunto de dados tipicamente explorado em tarefas de descoberta de conhecimento baseados em técnicas de clustering.

Entre outros objetivos, a integração entre clustering e CBR pode focalizar a utilização de algoritmos de clustering na escolha de quais atributos devem ser utilizados na indexação de casos. Em situações em que casos são capturados por listas muito grandes de atributos e valores, identificar os atributos mais relevantes para a solução de problemas de aplicação não é uma tarefa de indexação simples de ser realizada. Se atributos pouco representativos desses problemas são escolhidos para serem os índices desses casos, tarefas de recuperação de casos em sistemas CBR podem não alcançar resultados satisfatórios. Esta dissertação, neste caso, focaliza a exploração da integração entre clustering e CBR no apoio ao processo de identificação de atributos relevantes para a solução de problemas de CBR. Para expressar a relevância desses atributos, assim como explorado no nosso projeto, a significância de atributos é capturada pela utilização de pesos, os quais são associados a cada atributo utilizado nas computações de similaridade entre casos. De acordo com o processo de análise de índices para sistemas CBR proposto nesta dissertação, é fundamental observar que a análise da relevância de atributos via algoritmos de clustering somente pode acontecer se a função de similaridade utilizada em CBR for a mesma função usada pelos algoritmos de clustering explorados.

Na investigação de índices para a recuperação e organização de bases de casos de sistemas CBR, três tipos diferentes de algoritmos de clustering foram selecionados e explorados nesta dissertação: algoritmo baseado em densidade, algoritmo hierárquico e algoritmo particional (JAIN, ANIL K., 2008; SNEATH; SOKAL, 1973). Baseado na ideia de densidade de dados, o algoritmo DBScan foi escolhido por proporcionar a representação de grupos de casos que possuem formas arbitrárias. Algoritmos do tipo hierárquico são comumente

explorados na investigação de taxonomias (SNEATH; SOKAL, 1973) a partir da análise sistemática de dados, assim o algoritmo hierárquico DIANA foi escolhido por possibilitar uma visualização hierárquica de grupos de casos. Permitindo criar determinada quantidade de grupos a partir de dados usados como entrada, o algoritmo particional K-Means foi escolhido por ser um dos algoritmos mais tradicionais quando se requer um número exato de grupos.

Para cada algoritmo de clustering utilizado no processo de indexação de casos para CBR proposto nesta dissertação, parâmetros de entrada para esses algoritmos devem ser escolhidos e explorados. No algoritmo de densidade, a *quantidade mínima de casos* (minpts) que deve existir em um espaço para formar um grupo de casos e o *raio* (eps) que um grupo de casos deve abranger. No algoritmo hierárquico, embora não seja necessário informar parâmetros em muitas implementações, é relevante definir o *critério de ligação* (linkage) entre os grupos de casos. Por fim, no algoritmo particional, é necessário informar o *número de grupos* (k) a ser criado.

Tomando como entrada os casos disponíveis em uma base de casos, assim como testado no nosso projeto, a execução do algoritmo hierárquico deve ocorrer antes da execução dos demais algoritmos de clustering. Na prática, definiu-se por explorar um nível de corte em uma árvore, ou dendrograma, resultante da execução do algoritmo hierárquico. Isso permite definir um número de grupos de casos k a ser utilizado como entrada no algoritmo particional. Tal definição também é relevante na análise da quantidade mínima de pontos (minpts) como entrada no algoritmo de densidade, visto que o objetivo (para fins de comparação entre os diferentes algoritmos de clustering) é explorar o mesmo número de grupos de casos assim como obtido pelos diferentes algoritmos de clustering utilizados. De modo diferente ao particional, o algoritmo de densidade não define a quantidade de grupos que se quer formar. Entretanto, define-se o mínimo de casos que um grupo necessita ter para a sua formação. Portanto, conforme casos são agrupados no algoritmo de densidade, grupos de casos são formados até que todos os casos sejam agrupados, resultando em um total de grupos.

3.2 PROCESSO DE ANÁLISE DE ÍNDICES PARA CBR BASEADO NA AVALIAÇÃO DE AGRUPAMENTOS DE CASOS OBTIDOS A PARTIR DA EXECUÇÃO DE ALGORITMOS DE CLUSTERING

O processo de execução de algoritmos de clustering e avaliação de grupos de casos resultantes dessas execuções desenvolvido nesta dissertação para apoiar a investigação de pesos para atributos usados nas computações de similaridade entre casos em CBR foi organizado em três etapas/passos, conforme descrito no pseudocódigo do Algoritmo 1:

Passo 1) este passo utiliza um esquema de análise de índices onde “não existe uma distinção de relevância”, ou distinção de pesos, entre os atributos utilizados nas computações de similaridade entre casos realizadas tanto em CBR quando em clustering. Sendo assim, todos os atributos utilizados na indexação de casos para CBR têm uma “relevância idêntica” (mesmos pesos – no Algoritmo 1, ver o parâmetro EQUAL_WEIGHTING_SCHEME);

Passo 2) este passo utiliza um esquema de análise de índices onde existe uma avaliação de relevância, ou peso, a qual é “direcionada para cada atributo” usado nas computações de similaridade entre casos realizadas tanto em CBR quando em clustering. Sendo assim, a relevância de cada atributo utilizado na indexação de casos para CBR é avaliada “individualmente”. Na prática, um atributo vai ter um peso bastante alto em relação a um peso baixo e idêntico para os demais atributos utilizados nas computações de similaridade entre casos - no Algoritmo 1, ver os parâmetros SINGLE_WEIGHTING_SCHEME, onde um atributo é selecionado por caseAttributeIndex;

Passo 3) este passo utiliza um esquema onde a avaliação de relevância de todos os atributos usados nas computações de similaridade entre casos é “ajustada”, tanto nas computações realizadas em CBR quanto em clustering. Sendo assim, cada atributo utilizado na indexação de casos para CBR contém um peso associado à sua relevância relativa, assim resultando em uma função de similaridade ajustada à análise de relevância observada nos passos 1 e 2 - no Algoritmo 1, ver o parâmetro ADJUSTED_WEIGHTING_SCHEME.

Nesta dissertação, foram selecionados e explorados três diferentes métricas de avaliação de qualidade na avaliação de grupos de casos resultantes da execução dos algoritmos de clustering selecionados: a) entropia, b) precisão e c) pureza (no Algoritmo 1, ver o parâmetro CLUSTER_EVALUATION_METRIC). O objetivo de usar diferentes métricas de avaliação de qualidade associadas a estes algoritmos é permitir identificar qual dessas técnicas apresenta os melhores resultados para as necessidades de recuperação de casos na aplicação CBR sendo considerada. Além disso, ao utilizar diferentes técnicas, resultados semelhantes obtidos a partir do emprego de técnicas alternativas podem permitir melhor avaliar se esses resultados são consistentes entre si ou não. Em geral, resultados similares obtidos via técnicas distintas tendem a ser considerados mais robustos. Via métricas de entropia, precisão e pureza, o processo de avaliação de grupos de casos formados a partir da execução de diferentes algoritmos de

clustering, assim como proposto nesta dissertação, pode permitir que usuários analisem a homogeneidade dos grupos de casos formados.

Nesta dissertação, resultados obtidos a partir da execução dessas três métricas de avaliação de grupos de casos são explorados na construção de funções de similaridade ajustadas, as quais são utilizadas tanto em clustering quanto em CBR. Na prática, esses resultados permitem analisar o peso de cada atributo usado na representação de casos, os quais são indexados por uma função de similaridade ajustada para a resolução de problemas na aplicação alvo. No entanto, ajustar essas funções de similaridade diretamente com os resultados gerados por essas métricas de avaliações de qualidade de grupos de casos não é uma tarefa simples. Entre outros problemas, as escalas numéricas dos valores resultantes da utilização dessas métricas podem não ser coerentes entre si, por exemplo. Para contornar problemas encontrados, assim como explorado nesta dissertação, os resultados dessas métricas de avaliação de clustering foram normalizados entre os valores 1.00 e 10.00. Entre outros motivos, essa escala de valores é simples de ser analisada por usuários e, conforme testes realizados no nosso projeto, os valores representados nessa escala permitem visualizar a relevância de atributos utilizados nas computações de similaridade realizadas em CBR e clustering. Em resumo, existe a necessidade de normalizar os valores obtidos a partir da utilização de métricas como pureza, entropia e precisão (no Algoritmo 1, ver o parâmetro NORMALIZATION_METHOD). Para isso, dois métodos de normalização foram explorados neste trabalho:

(A) Normalização linear:

O menor valor obtido em uma determinada métrica de avaliação de qualidade de grupos de casos é associado ao valor 1.00 (quando usado em uma função de similaridade, esse valor vai representar um peso = 1.00). Em contraste, o maior valor é associado ao valor 10.00 (esse valor vai representar um peso = 10.00). Demais valores são normalizados linearmente dentre esses dois extremos.

(B) Normalização logarítmica

O menor valor obtido em uma determinada métrica de avaliação de qualidade de grupos de casos é associado ao valor 1.00 (quando usado em uma função de similaridade, esse valor vai representar um peso = 1.00). Em contraste, o maior valor é associado ao valor 10.00 (esse valor vai representar um peso = 10.00). Demais valores obtidos são normalizados logaritmicamente dentre esses dois extremos. Em particular, logaritmos podem ser utilizados

pois valores baixos e próximos a 1.00 crescem mais rapidamente que valores altos. O uso desta técnica justifica-se pelo fato de que pesos formados linearmente com valores baixos e próximos a 1.00 podem não influenciar tanto na relevância de atributos utilizados nas computações de similaridade entre casos.

Algoritmo 1 – Algoritmo de execução de clustering e avaliação de grupos de casos formados.

Entrada: Base de casos: cbr; algoritmo de clustering: clusterAlgorithm; objetos representando grupos de casos formados: clusters

Saída: Conjunto de pesos para atributos indexados: cbr

Método

```
01  //(A) Exploração de um esquema de avaliação de relevância idêntico
02  cbr.weightValuesForCaseAttributes = cbr.setWeightValues( EQUAL_WEIGHTING_SCHEME );
03  clusters.equalWeightingResults = clusterAlgorithm.execute( cbr.getCaseBase(), cbr.weightValuesForCaseAttributes );
04  clusters.equalWeightingCentroids = clusters.centroidsFromClusterResults( clusters.equalWeightingResults );
05  clusters.equalWeightingEvaluation = clusters.evaluateClusterResults(clusters.equalWeightingResults, CLUSTER_EVALUATION_METRIC );
06  //(B) Exploração de um esquema de avaliação de relevância individual
07  for caseAttributeIndex = 0 to cbr.getNumberOfCaseAttributes() do
08          cbr.weightValuesForCaseAttributes = cbr.setWeightValues( SINGLE_WEIGHTING_SCHEME, caseAttributeIndex );
09          clusters.singleWeightingResults[caseAttributeIndex] =
                    clusterAlgorithm.execute( cbr.getCaseBase(), cbr.weightValuesForCaseAttributes );
10          clusters.singleWeightingEvaluations[caseAttributeIndex] =
                    clusters.evaluateClusterResults(clusters.singleWeightingResults[caseAttributeIndex], CLUSTER_EVALUATION_METRIC );
11  Endfor
12  //(C) Exploração de um esquema de avaliação de relevância ajustado
13  cbr.weightValuesForCaseAttributes = cbr.setWeightValues(
            ADJUSTED_WEIGHTING_SCHEME,
            clusters.normalizeClusterEvaluationsIntoWeightValues( clusters.singleWeightingEvaluations, NORMALIZATION_METHOD ) );
14  clusters.adjustedWeightingResults = clusterAlgorithm.execute( cbr.getCaseBase(), cbr.weightValuesForCaseAttributes );
15  clusters.adjustedWeightingCentroids = clusters.centroidsFromClusterResults( clusters.adjustedWeightingResults );
16  clusters.adjustedWeightingEvaluation =
            clusters.evaluateClusterResults( clusters.adjustedWeightingResults, CLUSTER_EVALUATION_METRIC);
```

3.3 EXPLORAÇÃO DE UM ESQUEMA DE AVALIAÇÃO DE RELEVÂNCIA "IDÊNTICO" NA ANÁLISE DE ÍNDICES PARA SISTEMAS CBR

Para analisar a qualidade de grupos de casos obtidos quando algoritmos de clustering são executados, é relevante analisar essa qualidade em relação a valores básicos utilizados como referência. Sendo assim, a ideia é obter resultados tomados como baseline que possam ser comparados com outros resultados de avaliação de grupos de casos, apoiando a identificação de resultados possivelmente não satisfatórios.

Resultados utilizados como baseline são obtidos quando um mesmo peso é associado a todos os atributos usados nas computações de similaridade entre casos. Valores baseline são computados para cada uma das métricas de avaliação de grupos de casos, onde tais grupos são resultantes da execução de algoritmos de clustering. Neste caso, i) os algoritmos de clustering são executados, onde todos os atributos utilizados nas computações de similaridade entre casos são associados ao peso 1.00 (Algoritmo 1 – linha 2, onde a constante EQUAL_WEIGHTING_SCHEME define o peso 1.00). Na prática, é utilizada uma função de similaridade onde todos os atributos utilizados têm o mesmo peso; ii) os centroides obtidos com esta execução de clustering são armazenados para apoiar na futura determinação de sub-bases de casos (Algoritmo 1 – linha 4), as quais podem ser utilizadas na computação de respostas para consultas CBR (ver estudo de caso detalhado no Capítulo 4); e iii) as métricas de avaliação dos grupos de casos são executadas e esses resultados baseline são armazenados para apoiar futuras análises (Algoritmo 1 – linha 5). Em resumo, cada algoritmo de clustering e métrica executados no nosso processo de análise de índices devem ter um valor baseline correspondente.

3.4 EXPLORAÇÃO DE UM ESQUEMA DE AVALIAÇÃO DE RELEVÂNCIA "INDIVIDUAL" NA ANÁLISE DE ÍNDICES PARA SISTEMAS CBR

Neste passo, o objetivo é criar e avaliar grupos de casos a partir do ajuste dos pesos de cada atributo usado nas computações de similaridade entre casos (Algoritmo 1 – linha 7). Para analisar a relevância de cada um destes atributos, as métricas entropia, pureza e precisão são executadas sobre os grupos de casos obtidos quando os algoritmos de clustering são executados. Na prática, um valor "alto" de peso é atribuído para cada atributo utilizado na representação de casos (Algoritmo 1 – linha 8, onde a constante SINGLE_WEIGHTING_SCHEME define um valor de peso alto, tal como $peso_{atr_i} = 100.00$). Além disso, os demais atributos usados na representação dos casos são associados a um peso = 1.00. Em resumo, a utilização deste valor

de peso “alto” nestes atributos busca realçar o impacto de cada atributo nas computações de similaridade entre casos. Para tais computações de similaridade, uma equação de similaridade Euclidiana ponderada pode ser utilizada, assim como explorado neste trabalho, visto que essa medida é simples e genérica. No entanto, outras medidas de similaridade também podem ser exploradas quando necessário.

Usando uma configuração de pesos nos atributos que indexam os casos armazenados na base de casos, um algoritmo de clustering é utilizado para obter grupos de casos (Algoritmo 1 – linha 09). Para cada atributo utilizado nas computações de similaridade, e utilizando cada uma das métricas de avaliação de qualidade selecionadas, diferentes valores de avaliação de grupos de casos são obtidos (Algoritmo 1 – linha 10, onde a constante CLUSTER_EVALUATION_METRIC define a métrica a ser utilizada). Na prática, esses valores de avaliação de qualidade de grupos permitem observar o impacto de cada um dos atributos na formação de grupos de casos mais ou menos homogêneos no problema de aplicação sendo considerado.

3.5 EXPLORAÇÃO DE UM ESQUEMA DE AVALIAÇÃO DE RELEVÂNCIA “AJUSTADO” NA ANÁLISE DE ÍNDICES PARA SISTEMAS CBR

Uma vez que resultados da execução das métricas de avaliação de grupos produzidos são obtidos, eles devem ser normalizados buscando obter estimativas de relevância para atributos utilizados nas computações de similaridade entre casos. Para isso, diferentes métodos de normalização podem ser explorados (Algoritmo 1 – linha 13, onde o parâmetro NORMALIZATION_METHOD define o tipo de normalização a ser utilizado). Neste trabalho, esquemas de normalização alternativos foram explorados:

(A1) Normalização linear de todos os valores resultantes da execução das métricas de avaliação de grupos de casos

Valores resultantes das métricas de avaliação de grupos de casos, os quais são obtidos quando os algoritmos de clustering são executados, são normalizados linearmente entre 1.00 e 10.00. Neste processo, resultados tomados como baseline não são considerados na seleção de um subconjunto de valores resultantes destas métricas a ser normalizados. Neste caso, todos os atributos são ponderados, mesmo os possivelmente irrelevantes associados a valores de avaliação de qualidade inferiores ao valor baseline, assim como obtido no passo 1 do processo de investigação de índices proposto;

(A2) Normalização linear apenas dos valores resultantes da execução das métricas de avaliação de grupos de casos que são maiores que valores tomados como baseline

Valores de baseline são considerados na seleção dos valores das métricas que devem ser normalizados, enquanto valores resultantes da execução das métricas de avaliação de grupos de casos que são menores que valores tomados como baseline são normalizados para o valor 1.00. Neste caso, avaliações numéricas de qualidade de grupos de casos iguais ou abaixo de valores tomados como baseline são normalizadas para o valor 1.00. Em contrapartida, os valores acima do baseline são normalizados entre um mínimo e um máximo valor resultante da execução das métricas de avaliação de grupos de casos, os quais devem ser maiores que o valor tomado como baseline;

(B1) Normalização logarítmica de todos os valores resultantes da execução das métricas de avaliação de grupos de casos

Valores resultantes das métricas são normalizados logaritmicamente entre 1.00 e 10.00, sem considerar os resultados de baseline. Quando este método de normalização é utilizado, os valores superiores e próximos a 1.00 crescem mais rapidamente que valores próximos a 10.00. Neste caso, todos os atributos utilizados nas computações de similaridade entre casos são ponderados, assim buscando melhor expressar o impacto de valores baixos de peso nas computações de similaridade;

(B2) Normalização logarítmica apenas dos valores resultantes da execução das métricas de avaliação de grupos de casos que são maiores que valores tomados como baseline

Neste método, valores de baseline são considerados. Neste caso, resultados de avaliações de qualidade de grupos de casos iguais ou abaixo de valores tomados como baseline são associados ao valor 1.00. Os valores acima do baseline são normalizados entre um mínimo e um máximo valor, os quais devem ser maiores que o valor tomado como baseline;

Com as estimativas de relevância (ou pesos) de cada um dos atributos usados nas computações de similaridade entre casos, as quais são obtidas a partir da normalização dos valores resultantes de computações de entropia, pureza e precisão sobre os grupos de casos obtidos quando os algoritmos de clustering são executados, deve-se então empregar cada um desses conjuntos de pesos na função de similaridade utilizada tanto pelos algoritmos de clustering quanto pelos algoritmos de CBR. O objetivo é avaliar se tais pesos associados aos

atributos usados nas computações de similaridade entre casos conseguem apoiar na formação de grupos de casos mais ou menos homogêneos na aplicação CBR sendo considerada. Para isso, a função de similaridade é ajustada com as estimativas de peso obtidas para cada um dos atributos usados nas computações de similaridade, assim como detalhado no passo 2 do nosso método de análise de índices para CBR.

Na prática, os i) grupos de casos formados a partir da execução dos algoritmos de densidade, hierárquico e particional utilizando uma função de similaridade ajustada podem ser diferentes dos ii) grupos de casos formados pela utilização de uma função de similaridade onde todos os valores de peso dos atributos utilizados nos cálculos de similaridade entre casos sejam idênticos. Mais ainda, esta tarefa de avaliação de grupos de casos pode ser realizada utilizando diferentes estimativas numéricas de peso para estes atributos. Estas diferentes estimativas podem ser obtidas quando diferentes métodos de normalização são utilizados, assim como descrito anteriormente. Portanto, o processo de grupo (Algoritmo 1 – linha 14) é executado novamente, onde os grupos resultantes são armazenados para futuras avaliações. Ainda, os centroides de cada grupo obtido são também armazenados (Algoritmo 1 – linha 15), assim permitindo usar esses centroides como índices para sub-bases de casos e consequente computação de respostas para consultas CBR. Neste caso, os grupos aqui formados são baseados na utilização de uma função de similaridade ajustada com diferentes estimativas de pesos para os atributos utilizados nessa função. No final, resultados de avaliação de qualidade desses grupos são obtidos novamente (Algoritmo 1 – linha 16, onde a constante CLUSTER_EVALUATION_METRIC define a métrica), visando permitir a análise da qualidade desses grupos em relação as necessidades de indexação sendo tratadas no desenvolvimento de sistemas CBR.

3.6 AVALIANDO A ACURÁCIA DE CONSULTAS CBR A PARTIR DA UTILIZAÇÃO DOS ÍNDICES OBTIDOS QUANDO A ABORDAGEM DE CLUSTERING É UTILIZADA

Para analisar a qualidade dos resultados obtidos no processo de indexação de casos descrito anteriormente, o método cross-validation pode ser utilizado no teste dos algoritmos de CBR. Nestes testes, um caso da base de casos é selecionado e removido desta base. Em seguida, esse caso é utilizado como consulta no sistema CBR. Uma vez que cada uma dessas consultas tenha sido executada, e o resultado de cada uma tenha sido anotado (se o sistema apresentou

uma solução aceitável ou não para o problema descrito na consulta), a acurácia final do sistema CBR pode ser computada.

De acordo com trabalhos relacionados (MITTAL et al., 2014; VERNET; GOLOBARDES, 2003; YANG; WU, 2000), utilizar sub-bases de casos como índices na computação de resultados para consultas CBR pode melhorar a acurácia destas consultas. Quando essas sub-bases de casos são usadas como índices na recuperação de casos em CBR, um primeiro passo visa computar a similaridade de um caso tomado como consulta em relação a um "caso prototípico", ou centroide, que representa as características médias dos casos contidos em uma sub-bases de casos selecionada. Essa avaliação de similaridade entre consultas e sub-bases de casos é repetida em cada uma das sub-bases de casos definidas na memória de um sistema CBR. Em um segundo passo, novas avaliações de similaridade são computadas somente considerando os casos que estão presentes naquela sub-base de casos que é a mais similar a consulta dada. A partir desses dois passos, respostas para consultas CBR podem ser obtidas.

A ideia de construir e utilizar sub-bases de casos como índices para a computação de consultas CBR foi explorada neste projeto. Para isso, tais sub-bases de casos foram formadas a partir dos resultados de grupos de casos formados a partir das várias execuções dos algoritmos de clustering. Mais ainda, a definição dessas sub-bases de casos pode ser realizada de diferentes formas. Neste caso, grupos de casos tomados como sub-bases de casos podem ser oriundos da execução de diferentes algoritmos de clustering. Na prática, diferentes grupos de casos podem ser obtidos quando um algoritmo de clustering é executado múltiplas vezes (por exemplo, para cada conjunto de pontos aleatórios tomados inicialmente, o algoritmo particional pode produzir diferentes grupos), ou quando diferentes algoritmos de clustering são executados. Além disso, ajustes nas funções de similaridade utilizadas por algoritmos de clustering também podem resultar em diferentes grupos de casos e sub-bases de casos correspondentes.

No nosso projeto, estimativas numéricas de peso associadas a cada um dos atributos utilizados nas computações de similaridade para a formação de grupos de casos foram obtidas e exploradas. Estes grupos foram então tomados como sub-bases de casos utilizadas na computação de respostas para consultas CBR. O objetivo foi testar se a resposta de consultas CBR utilizando i) sub-bases de casos criadas a partir da utilização de uma função de similaridade ajustada (Algoritmo 1 – linha 14), onde os atributos utilizados nessa função foram ponderados e os centroides dos grupos formados foram utilizados como índices dessas sub-bases de casos (Algoritmo 1 – linha 15), poderia resultar em melhores valores de acurácia em comparação a valores de acurácia obtidos quando ii) sub-bases de casos criadas a partir da

utilização de uma função de similaridade não ajustada (Algoritmo 1 – linha 3) foram utilizadas, onde todos os atributos utilizados nessa função têm igual relevância e os centroides também são utilizados como índices para sub-bases de casos (Algoritmo 1 – linha 4). Neste trabalho, essa análise foi explorada da seguinte forma: para cada algoritmo de clustering (densidade, hierárquico ou particional), a função de similaridade foi ajustada segundo os pesos definidos a partir da execução das métricas de entropia, precisão ou pureza. Em seguida, os grupos formados conforme o algoritmo de clustering e métrica de avaliação escolhidos foram usados como sub-bases de casos em consultas CBR. Dessa forma, foi possível comparar qual dos algoritmos de clustering e qual das métricas de avaliação de qualidade de grupos de casos permitiu formar sub-bases de casos onde melhores resultados de acurácia em consultas CBR foram obtidos, assim como descrito no estudo de caso detalhado no Capítulo seguinte dessa dissertação.

3.7 CONSIDERAÇÕES DO CAPÍTULO

A etapa de recuperação de casos em sistemas CBR possui adversidades relacionadas à descoberta da relevância de atributos de casos para serem usadas em computações de similaridade entre casos. Estruturas de índices podem ser usadas para resolver estas adversidades. Para isso, foi apresentado neste Capítulo um processo onde algoritmos de clustering são executados e grupos de casos resultantes são avaliados visando a investigação de índices que permitam melhorar a acurácia de mecanismos de recuperação de casos em sistemas CBR.

Neste Capítulo, a proposta apresentada indica que quando algoritmos de clustering são explorados, assim como descrito no processo de análise de índices proposto, é possível obter feedback para construir estruturas de índice para sistemas CBR. Para isso, é construída e analisada uma função de similaridade ajustada, onde um valor muito alto como peso (por exemplo, 100.00) para um atributo selecionado é utilizado, enquanto os demais atributos representados em um caso são associados a um valor baixo de peso (1.00). A partir dessa configuração na função de similaridade, busca-se identificar o impacto de atributos selecionados na formação de melhores grupos de casos. Neste cenário, o cálculo de similaridade é ajustado para reforçar as similaridades oriundas de cada um dos atributos utilizados nos cálculos de similaridade realizados pelos algoritmos de clustering. Desta forma, o resultado desse reforço é observado a partir da análise da qualidade dos grupos de casos obtidos quando esses algoritmos de clustering são executados. Assim como explorado nesta dissertação, estas

estimativas de qualidade então podem indicar a relevância de cada atributo. Mais ainda, grupos de casos formados em clustering também podem ser explorados na definição de sub-bases de casos a serem utilizadas como índices para a resposta de consultas CBR. Em geral, o objetivo é organizar a base de casos para que computações de similaridade sejam realizadas em grupos que contenham casos mais semelhantes ao caso usado como consulta, ao invés destas computações serem realizadas em relação a todos os casos da base de casos.

Por fim, cabe ressaltar que os resultados obtidos a partir da exploração de algoritmos de clustering por si só são de interesse de usuários finais em diferentes domínios de aplicação. Em muitos sentidos, esses resultados proporcionam a descoberta de inferências a serem investigadas e exploradas nestes problemas de aplicação.

4 UMA AVALIAÇÃO DO FRAMEWORK PARA A INTEGRAÇÃO ENTRE CLUSTERING E CBR EM UM ESTUDO DE CASO ENVOLVENDO UM PROBLEMA DE SIMULAÇÃO

Este capítulo descreve uma avaliação realizada neste projeto de pesquisa para analisar o framework de integração entre clustering e CBR proposto. Para realizar esta avaliação, foi utilizado o framework para a investigação de índices para a recuperação e organização de casos em bases de casos de sistemas CBR assim como descrito no Capítulo 3. Nesta avaliação, um estudo de caso foi desenvolvido no contexto de um projeto de pesquisa envolvendo o projeto e prototipação de um sistema de simulação virtual tática para o Exército Brasileiro – o projeto SIS-ASTROS. Além disso, avaliações utilizando bases de casos da web (LICHMAN, 2013) também foram realizadas. Nestes problemas de aplicação, resultados de clustering e de consultas CBR são descritos e discutidos.

4.1 ESTUDO DE CASO

O projeto SIS-ASTROS busca pesquisar e prototipar um sistema integrado para a simulação virtual tática de tarefas relativas ao reconhecimento, escolha e ocupação de posições para baterias de artilharia. Nesse sistema, diferentes tipos de comportamentos automatizados são realizados por agentes simulados durante os exercícios de simulação. Na investigação e implementação desses comportamentos, esta dissertação desenvolveu um estudo piloto abordando um problema relacionado à escolha de suprimentos a serem utilizados por unidades autônomas simuladas. Baseado em recursos de CBR, o objetivo é realizar a correta escolha de suprimentos de forma a permitir que uma determinada missão de bateria de artilharia seja finalizada com sucesso.

A escolha de suprimentos a serem utilizados em missões de bateria de artilharia é um problema complexo a ser simulado em um nível de representação fiel a realidade. Além de fazer considerações sobre tipos disponíveis e quantidades de suprimentos de baterias de artilharia, a solução para tal problema certamente envolve a avaliação de fatores específicos de domínio e do contexto de uma batalha sendo simulada. Neste caso, a modelagem e implementação de algoritmos inteligentes para apoiar a resolução deste problema de simulação está além da exploração de um grande número de regras para representar conhecimento específico de domínio de uma doutrina militar. Em muitos sentidos, isso indica que soluções para tais problemas de seleção de suprimentos são alcançadas quando avaliações caso a caso são

desenvolvidas por militares altamente treinados. Em IA, CBR é uma técnica central para a modelagem de comportamentos de decisão para problemas como estes.

4.2 EXPERIMENTO REALIZADO UTILIZANDO A BASE DE CASOS DO PROJETO SIS-ASTROS

Para avaliar o framework proposto nesta dissertação, uma série de atividades foram realizadas assim como demonstrado na Figura 2. As atividades exploradas são: a) análise da base de casos e realização de implementações preliminares para a configuração do framework; b) configuração dos algoritmos de clustering; c) testes feitos em clustering segundo o esquema de avaliação de relevância “idêntico”; d) testes feitos em clustering segundo o esquema de avaliação de relevância “individual”; e) testes feitos em clustering segundo o esquema de avaliação de relevância “ajustado”; f) configuração da aplicação CBR; g) testes feitos em CBR utilizando pesos obtidos com o esquema de relevância “idêntico”; h) testes feitos em CBR utilizando pesos obtidos com o esquema de relevância “ajustado”; i) testes feitos em CBR utilizando pesos obtidos com o esquema de relevância “idêntico” e utilizando sub-bases de casos; e, j) testes feitos em CBR utilizando pesos obtidos com o esquema de relevância “ajustado” e utilizando sub-bases de casos.

4.2.1 Base de casos e implementações preliminares

Neste trabalho, o objetivo da aplicação de simulação explorada é construir um componente inteligente baseado em casos, o qual é associado a agentes envolvidos em simulações. Este componente deve permitir responder, de forma exata, consultas por casos similares no contexto da solução de problemas de simulação envolvendo a escolha de suprimentos a serem usados por agentes simulados. Desta forma, o framework proposto nesta dissertação foi utilizado para apoiar a investigação de índices para a recuperação e organização da base de casos utilizados por esse componente de simulação inteligente.

A base de casos utilizada neste estudo de caso é originária da investigação desenvolvida por JUNIOR (2016). Nesta base de casos, casos são representados por 13 atributos do problema de aplicação. A solução deste problema, assim como representada no corpo desses casos, é composta por 3 atributos. Sendo baseada nestes 3 atributos, tal solução está em contraste com aplicações de CBR onde uma solução é composta por apenas um atributo representando um “rótulo” de classe. Assim como observado nesta dissertação, o fato de uma solução ser

composta por 3 diferentes atributos apresenta complexidades adicionais no que diz respeito à avaliação de qualidade de grupos de casos formados quando algoritmos de clustering são executados. Sendo assim, para avaliar a precisão de um grupo de casos, foi necessário avaliar se soluções recuperadas estavam corretas de acordo com soluções presentes no corpo de casos utilizados como consulta. Para isso, foi necessário implementar uma função de decisão específica para este problema de aplicação, assim como proposta em JUNIOR (2016). Na aplicação de simulação utilizada neste estudo de caso, 3 atributos foram utilizados na descrição de soluções para os problemas de escolha de suprimentos. Conforme os 3 atributos que compõem a solução destes casos, uma função de similaridade entre atributos utilizados na representação de soluções em casos para CBR foi implementada nesta dissertação. Na prática, essa função foi utilizada para avaliar se casos foram devidamente agrupados ou não.

Para atacar o problema de simulação citado, implementações preliminares foram realizadas visando satisfazer regras específicas de domínio: a) implementação de funções de similaridade locais deste domínio de aplicação, as quais foram ajustadas para os atributos utilizados na representação de casos neste problema de aplicação, assim como propostas por JUNIOR (2016); b) implementação de funções de avaliação de qualidade de grupos de dados, neste caso: entropia, precisão e pureza, as quais foram ajustadas para medir a qualidade de grupos formados de acordo com regras de domínio existentes neste problema de aplicação.

Em geral, para realizar uma medida de precisão em clustering, por exemplo, um caso de um grupo de casos é removido. Em seguida, a solução deste caso é comparada com as soluções contidas nos demais casos contidos neste grupo. Para realizar tal comparação, a função de decisão típica desta aplicação foi utilizada. Por exemplo, remove-se o caso C_a de um grupo e para cada caso vizinho C_i pertencente a este grupo, compara-se a solução de C_a com a solução de C_i. Para isso, a função de decisão avalia se os 3 atributos de ambas as soluções estão de acordo com tal função, definindo se o caso C_i está correto em relação a C_a. O caso C_a que já foi avaliado em relação aos demais não é avaliado novamente. A função de precisão utilizada é baseada na Equação 4 (Seção 2.2.5.1).

Para avaliar a pureza de um grupo de casos, é necessário definir qual é a máxima ocorrência de classes nos grupos obtidos quando algoritmos de clustering são executados. Para definir qual é a maior ocorrência de uma solução neste problema de aplicação, dois atributos utilizados para representar soluções nestes casos foram utilizados. Dessa forma, a solução de um caso é comparada com cada vizinho em um determinado grupo, avaliando quantitativamente quais são os maiores valores de atributos entre os casos comparados. Em resumo, a pureza dos grupos de casos obtidos foi analisada de acordo com esses dois atributos

de solução selecionados. Por exemplo, remove-se o caso C_a de um grupo, e para cada caso vizinho C_i pertencente a este grupo, compara-se a solução de C_a com a solução de $C_{i.}$ Para isso, a função de decisão avalia qual é a maior solução (quantitativamente em relação aos 2 atributos verificados), definindo se o caso C_i é maior que C_a. O caso C_a que já foi avaliado em relação aos demais não é avaliado novamente. A função de pureza utilizada é baseada na Equação 6 (Seção 2.2.5.3).

Para avaliar a entropia de um grupo de casos, é necessário avaliar a desordem nos grupos formados em relação aos atributos da solução dos casos. Essa desordem é calculada computando a probabilidade da ocorrência de um atributo do caso, em relação a quantidade de atributos existente em tal solução. Essa métrica avalia se casos são similares aos seus vizinhos, assim buscando medir a homogeneidade (desordem) dentro de um grupo. Por exemplo, conta-se todas as possibilidades existentes para cada atributo^i da solução de casos em um determinado grupo, e, em seguida, computa-se a entropia para cada atributo da solução, onde calcula-se a quantidade de possibilidades de cada atributo ocorrer em relação ao número de atributos que compõe a solução, multiplicando pelo logaritmo desse mesmo cálculo. A função de entropia utilizada é baseada na Equação 5 (Seção 2.2.5.2).

4.2.2 Configuração dos algoritmos de clustering

O objetivo deste passo é configurar os algoritmos de clustering para determinar um número de grupos formados por estes algoritmos que atenda aos objetivos da aplicação alvo, que seja padrão e utilizado em todos os algoritmos de clustering explorados, assim permitindo comparar a qualidade de grupos de casos resultantes da execução desses diferentes algoritmos de clustering. Para utilizar os algoritmos de clustering selecionados no processo de indexação de casos para CBR, diversas execuções destes algoritmos de clustering foram então realizadas. Em cada uma dessas execuções, diferentes combinações de parâmetros de entrada desses algoritmos foram exploradas, assim buscando encontrar valores para estes parâmetros que resultassem em grupos de casos ajustados para o problema de aplicação sendo tratado no estudo de caso.

(A) Exploração de parâmetros do algoritmo hierárquico

Este algoritmo utiliza como parâmetros de entrada o critério de linkage a ser utilizado. Este critério define a forma que o algoritmo deve medir a similaridade entre grupos de casos. Neste estudo de caso, os 3 critérios de linkage mais comumente utilizados por algoritmos

hierárquicos foram utilizados: Average, Complete e Single. Além disso, neste algoritmo, quando os grupos são formados, é possível definir um nível de corte na árvore representando os grupos obtidos a partir da execução deste algoritmo, assim permitindo selecionar uma determinada quantidade de grupos a ser usada em um problema de aplicação. Para cada critério de linkage utilizado, avaliações de qualidade nos grupos obtidos foram realizadas e estimativas de relevâncias dos atributos foram formadas. Em seguida, essas estimativas foram entradas nas funções de similaridade de clustering e CBR, validando a configuração realizada no algoritmo de clustering. De acordo com a definição de uma certa quantidade de grupos obtidos neste algoritmo, a partir do nível de corte na árvore, nos outros algoritmos o objetivo é encontrar configurações que resultem em uma quantidade de grupos similar a este algoritmo.

(B) Exploração de parâmetros do algoritmo de densidade

Este algoritmo utiliza como parâmetros de entrada a quantidade mínima de casos minpts para formar um grupo e o raio eps de abrangência mínimo para a formação de um grupo. Este algoritmo não permite definir um número de grupos a ser formado. No estudo de caso realizado, testes foram realizados testando diferentes valores de eps e minpts, até que este algoritmo produzisse em torno de 15 grupos como resultado de grupo de casos, assim como nos demais algoritmos. Nestes, por exemplo, para buscar algo em torno de 15 grupos, eps foi variado entre eps = 0.01 e eps = 20.00. Minpts foi variado entre minpts = 1 e minpts = 50. No estudo de caso tratado nesta dissertação, é relevante notar que os diferentes grupos obtidos com outros valores de eps e minpts produziram resultados de avaliações de qualidade similares aos resultados obtidos quando os parâmetros foram configurados para formar em torno de 15 grupos.

(C) Exploração de parâmetros do algoritmo particional

Este algoritmo utiliza como parâmetro de entrada a quantidade k de grupos a ser formada. Dessa forma, esse parâmetro foi ajustado em k = 15, como nos demais algoritmos. No entanto, testes adicionais foram realizados tentando descobrir um valor adequado para k. Por exemplo, nos testes realizados, k foi variado entre k = 2 e k = 50. Os grupos formados conforme esses outros valores para k, produziram resultados de avaliações de qualidade similares aos resultados obtidos quando esse parâmetro foi configurado para formar 15 grupos.

Os grupos formados pelos diferentes algoritmos de clustering foram analisados, permitindo verificar se eles atendiam as necessidades de identificação de grupos de casos existente no problema de aplicação. As análises realizadas demonstraram que essas formações

de grupos separavam os casos de acordo com atributos realmente significativos para o problema de aplicação. As análises visuais dos dendrogramas representando os grupos formados também demonstraram que foi possível obter grupos homogêneos com estas configurações de parâmetros de entrada. Contudo, outras explorações a respeito dos valores destes parâmetros de entrada ainda poderiam ser realizadas. No entanto, os valores alcançados com estes testes foram definidos com o objetivo de utilizar o mesmo número de grupos de casos na comparação dos resultados de avaliação associados aos diferentes algoritmos de clustering utilizados.

4.2.3 Testes feitos em clustering segundo o esquema de avaliação de relevância "idêntico"

O objetivo deste passo é obter valores baseline que possam ser utilizados no processo de análise da relevância de atributos utilizados em computações de similaridade entre casos. Neste passo, a função de similaridade utilizada pelos algoritmos de clustering selecionados é configurada de forma que cada atributo utilizado nesta função tenha um peso idêntico e igual a 1.00. Ao executar os algoritmos de clustering com essa função de similaridade tradicional, os grupos de casos resultantes são avaliados quanto a entropia, precisão e pureza. Nesse processo de avaliação, os resultados obtidos são armazenados como valores baseline, os quais são utilizados mais tarde no processo de normalização de outras estimativas de qualidade relacionadas a grupos de casos obtidos.

Neste passo, os grupos de casos formados são também armazenados em memória visando a utilização desses grupos em outras tarefas. Assim como explorado nesta dissertação, esses grupos são usados no processo de recuperação de casos em CBR. Para isso, o algoritmo de recuperação de casos utiliza os grupos de casos tomados como baseline como sub-bases de casos, as quais são utilizadas na resposta de consultas em sistemas CBR. Neste passo do processo de análise de índices para sistemas CBR, os diferentes resultados baseline obtidos podem ser comparados entre si, visto que diferentes algoritmos de clustering são explorados no desenvolvimento desta dissertação.

4.2.4 Testes feitos em clustering segundo o esquema de avaliação de relevância "individual"

O objetivo deste passo é investigar a relevância individual de cada atributo, ou peso do atributo, utilizado nas computações de similaridade entre casos. Para isso, diferentes

configurações de pesos são exploradas na função de similaridade utilizada pelos algoritmos de clustering, onde cada configuração focaliza cada um dos atributos utilizados por essa função. Neste caso, um atributo é selecionado e um peso = 100.00 é associado a ele, enquanto os demais atributos utilizados nessa função são associados a pesos idênticos e baixos, tal como 1.00. Ao executar os algoritmos de clustering com esta configuração de função de similaridade e obter uma avaliação dos grupos de casos formados, é possível observar a influência de cada um destes atributos na formação de grupos mais ou menos homogêneos. Para avaliar essa homogeneidade, os valores de avaliação destes grupos de casos (neste caso, utilizando as métricas de pureza, entropia e precisão) podem ser comparados com valores baseline obtidos no passo anterior desse processo de análise de índices. Se valores de avaliação de qualidade de grupos de casos obtidos para um desses atributos altamente ponderados na função de similaridade forem maiores que valores de avaliação de qualidade baseline, isso pode indicar que o atributo selecionado tem uma relevância alta (um peso alto em relação a outros pesos utilizados por outros atributos) nas computações de similaridade entre casos no domínio de aplicação considerado.

Neste passo, os resultados destas diferentes avaliações de qualidade de grupos de casos são armazenados em memória. Para um mesmo algoritmo de clustering, resultados obtidos para entropia, precisão e pureza então podem ser comparados entre si. Por exemplo, um resultado de entropia relacionado a um grupo de casos que reflete a alta ponderação do atributo 1 de um caso (um peso alto associado a este atributo na função de similaridade utilizada no processo de clustering) pode ser comparado ao resultado de entropia para o atributo 2, e assim por diante. Mais ainda, estes resultados de avaliação de qualidade de grupos de casos obtidos para cada atributo utilizado nas computações de similaridade podem ser comparados com os resultados de avaliação de qualidade tomados como baseline. Por fim, tomando uma mesma métrica de avaliação como referência, por exemplo: pureza, resultados obtidos quando essa métrica é medida podem permitir a comparação entre avaliações de grupos de casos obtidos quando diferentes algoritmos de clustering são executados. Neste caso, avaliações de pureza obtidas quando o algoritmo de clustering hierárquico é usado podem ser comparadas com avaliações de pureza obtidas quando outros algoritmos de clustering são utilizados. Por exemplo, um resultado de entropia para o atributo 1 obtido com o algoritmo de clustering hierárquico pode ser comparado com resultado de entropia para o atributo 1 obtido com o algoritmo de clustering particional. Comparações similares a estas podem ser realizadas para outros resultados obtidos.

4.2.5 Testes realizados em clustering segundo o esquema de avaliação de relevância "ajustado"

O objetivo deste passo é investigar se um dado conjunto de pesos associados aos atributos utilizados nas computações de similaridade executadas pelos algoritmos de clustering consegue ou não resultar na formação de grupos de casos mais ou menos homogêneos. Para obter essas estimativas de relevância, os resultados de avaliações de qualidade de grupos de casos obtidos quando o esquema de avaliação de relevância "individual" é executado são normalizados. Assim como explorado nesta dissertação, os métodos de normalização seguintes foram explorados: a) normalização linear sem considerar baseline; b) normalização linear considerando baseline; c) normalização logarítmica sem considerar baseline; e d) normalização logarítmica considerando baseline. Em seguida, a função de similaridade utilizada pelos algoritmos de clustering é configurada de forma que conjunto de pesos, os quais são definidos a partir desses métodos de normalização, sejam utilizados nas computações de similaridade entre casos.

Utilizando uma função de similaridade ajustada nos algoritmos de clustering utilizados, os resultados da avaliação de grupos de casos formados são também armazenados em memória. Isso permite comparar resultados de avaliação de qualidade de grupos de casos obtidos quando diferentes funções ajustadas de similaridade são utilizadas. Isso ocorre quando diferentes conjuntos de pesos para atributos utilizados por essas funções são obtidos quando os métodos de normalização citados são executados. Por exemplo, um resultado de entropia obtido com "normalização linear sem considerar baseline" pode ser comparado com um resultado de entropia obtido com "normalização linear considerando o baseline". De forma similar, outras comparações como esta podem ser realizadas.

Em resumo, o objetivo do processo de execução de algoritmos de clustering e avaliação de grupos de casos resultantes proposto nesta dissertação é encontrar uma configuração de índices a ser utilizada por algoritmos de CBR. Tais índices envolvem pesos para atributos usados em computações de similaridade entre casos, bem como o uso ou não de sub-bases de casos obtidas quando essas funções de similaridade ajustadas são usadas na execução desses algoritmos de clustering. No final, espera-se encontrar índices que permitem obter resultados de acurácia relevantes para consultas executadas em sistemas CBR.

4.2.6 Testes do sistema CBR

Em clustering e CBR, existe a necessidade de avaliar a qualidade de índices investigados. Tal avaliação de qualidade pode ser realizada no contexto da execução de algoritmos de clustering por si só, mas também pode ser realizada no contexto de execução de algoritmos de CBR. Em CBR, a acurácia de consultas pode ser avaliada de acordo com os diferentes conjuntos de pesos identificados para atributos utilizados nas computações de similaridade entre casos. Além disso, essa avaliação pode analisar a) resultados de consultas CBR que não utilizam as sub-bases de casos resultantes da execução dos algoritmos de clustering e b) resultados que utilizam essas sub-bases de casos.

Nesta dissertação, a avaliação de acurácia de consultas CBR é realizada em um processo de cross-validation. Para isso, todos os casos contidos na base de casos são explorados como consultas, e os resultados dessas consultas são utilizados na avaliação da acurácia do sistema. Para avaliar tal acurácia no estudo de caso realizado nesta dissertação, uma função de decisão utilizada pelos algoritmos de clustering é utilizada. Como os casos do problema de simulação tratado contêm soluções representadas por três diferentes atributos, essa função de decisão construída nesta dissertação para essa aplicação compara os valores destes atributos e informa se soluções entre pares de casos são realmente similares ou não (embora computações de similaridade entre soluções não sejam exploradas nas computações de similaridade entre casos realizadas em clustering e na recuperação de casos em CBR). A partir da utilização desse tipo de função de decisão, é possível avaliar se soluções contidas nos casos mais similares recuperados de uma base de casos para uma consulta dada são relevantes para resolver o problema encontrado no corpo da consulta.

Nos testes desenvolvidos nessa dissertação, esse processo de cross-validation é executado em diferentes formas.

A) Testes realizados em CBR utilizando pesos obtidos com o esquema de relevância “idêntico”

O objetivo deste passo é obter valores baseline de acurácia, os quais são explorados na análise de resultados de acurácia obtidos em outros testes realizados com os algoritmos de CBR. Nesse esquema de relevância “idêntico”, todos os atributos utilizados nas computações de similaridade têm o mesmo peso = 1.00. Em geral, esses valores baseline de acurácia associada aos algoritmos de CBR visam auxiliar na análise da qualidade das funções de similaridade ajustadas construídas, avaliando se os pesos para atributos obtidos a partir da execução do processo de análise de índices proposto nesta dissertação permitem obter respostas para consultas CBR com maior ou menor acurácia que respostas obtidas pelos algoritmos de

recuperação configurados conforme valores baseline. Em um processo de cross-validation associado a análise dos algoritmos de CBR, essas avaliações de acurácia obtidas para consultas CBR são então armazenadas em memória para permitir futura análise no processo de investigação de índices.

B) Testes realizados em CBR utilizando pesos obtidos com o esquema de relevância "ajustado"

O objetivo deste passo é descobrir se utilizar um conjunto de pesos formados, assim como obtidos de acordo com as métricas de avaliação de qualidade aplicados a resultados obtidos a partir da execução de algoritmos de clustering, conseguem apoiar consultas CBR a serem respondidas com maior ou menor acurácia. Para isso, conjunto de pesos são empregados na função de similaridade de CBR e avaliações de acurácia de consultas onde essa função ajustada é explorada são realizadas. Por exemplo, valores de peso oriundos da execução do algoritmo de densidade, onde avaliações de entropia de resultados desse algoritmo de clustering são normalizadas pela utilização do método linear sem considerar valores de baseline, são usados como entrada em uma função de similaridade ajustada. Utilizando essa função de similaridade ajustada a partir dessas estimativas de entropia normalizadas, o processo de cross-validation é executado na análise de consultas CBR. As avaliações de acurácia obtidas são então armazenadas em memória para permitir futura análise no processo de investigação de índices.

Neste passo, comparações entre os resultados de acurácia de consultas CBR obtidos quando cada uma das técnicas de formação de pesos é utilizada podem ser realizadas. Por exemplo, resultados de acurácia obtidos utilizando pesos formados com o algoritmo de densidade e a métrica "entropia", normalizados linearmente sem considerar o baseline, podem ser comparados com resultados de acurácia obtidos utilizando pesos formados com o algoritmo hierárquico e a métrica "entropia", e assim por diante. Mais ainda, resultados de acurácia utilizando funções de similaridade ajustadas nos algoritmos de CBR podem ser comparados a resultados de acurácia tomados como baseline (armazenados em memória, assim como descrito no passo A anterior).

C) Testes realizados em CBR utilizando pesos obtidos com o esquema de relevância "idêntico" e utilizando sub-bases de casos

O objetivo deste passo é descobrir se utilizar todos os atributos de casos com pesos idênticos (e iguais a 1.00) na função de similaridade utilizada por algoritmos de clustering na formação de sub-bases de casos pode resultar em melhores valores de acurácia associados a

análise de consultas CBR. A partir da execução destes testes, valores baseline de acurácia para consultas CBR que exploram sub-bases de casos são obtidos. Esses valores baseline são então explorados na análise de resultados de acurácia de consultas CBR obtidos quando sub-bases de casos são oriundas da utilização de um conjunto de pesos na função de similaridade utilizada por algoritmos de clustering na formação dessas sub-bases de casos.

Nas sub-bases de casos formadas segundo diferentes algoritmos de clustering, o centroide de cada um desses grupos de casos é rotulado como índice da sub-base de casos correspondente ao grupo. Com isso, a cada consulta CBR, o caso consultado tem a sua similaridade comparada em relação a cada índice (centroide) dessas sub-bases de casos. A partir disso, a sub-base de casos que possui o índice com a menor similaridade computada em relação ao caso consultado é selecionada. Em seguida, o caso consultado é comparado com cada caso pertencente a tal sub-base de casos selecionada. Por fim, os casos similares ao caso atual são então utilizados como resposta para a consulta CBR dada.

Neste passo, assim como em outros passos dos processos de análise de índices descritos nesta dissertação, os valores de baseline obtidos são armazenados em memória. Esses valores podem ser comparados a valores baseline obtidos quando diferentes algoritmos de clustering são explorados. Por exemplo, valores de acurácia obtidos utilizando sub-bases de casos formadas quando o algoritmo de densidade foi executado podem ser comparados com valores baseline obtidos utilizando sub-bases de casos formadas quando o algoritmo hierárquico foi executado, e assim por diante.

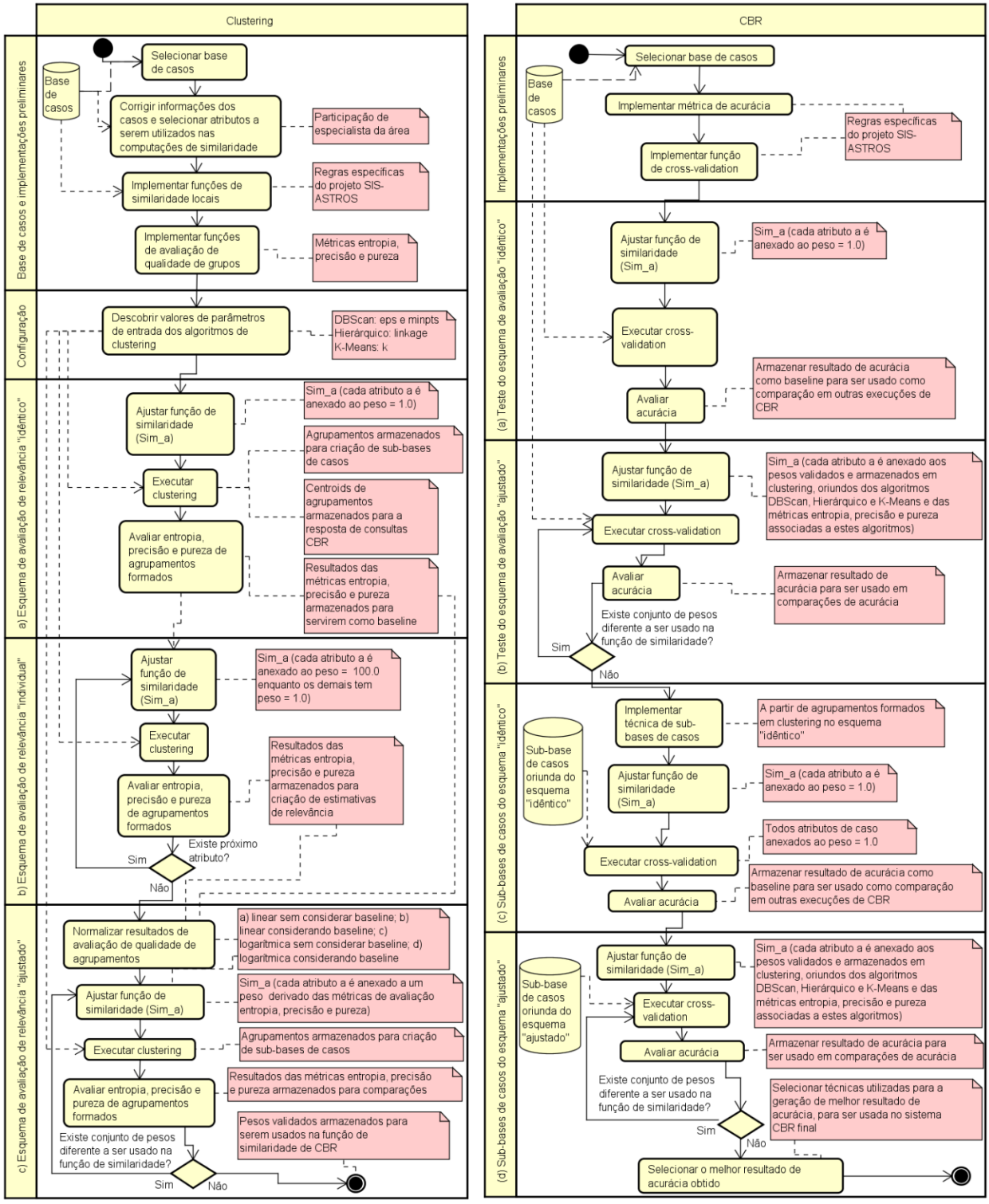
D) Testes realizados em CBR utilizando pesos obtidos com o esquema de relevância “ajustado” e utilizando sub-bases de casos

O objetivo deste passo é descobrir se a utilização de um conjunto de pesos na função de similaridade utilizada por algoritmos de clustering na formação de sub-bases de casos pode resultar em bons valores de acurácia associados a resolução de consultas CBR. A função de similaridade em clustering é então ajustada de acordo com um conjunto de pesos obtidos segundo alguma das técnicas de análise de índices exploradas. A partir desse ajuste, sub-bases de casos formadas são utilizadas na obtenção de respostas para consultas CBR.

Neste passo, os resultados de acurácia obtidos são armazenados em memória. Desta forma, cada resultado obtido pode ser comparado com resultados obtidos quando diferentes algoritmos de clustering são utilizados. Por exemplo, resultados de acurácia obtidos quando o mecanismo de recuperação de casos utiliza sub-bases de casos oriundas do algoritmo de clustering de densidade usando a função de similaridade ajustada com pesos formados a partir

da avaliação da métrica entropia, normalizados linearmente sem considerar o baseline, podem ser comparados com resultados de acurácia obtidos quando o mecanismo de recuperação de casos utiliza sub-bases de casos oriundas do algoritmo de clustering hierárquico com a função de similaridade ajustada com pesos formados a partir da avaliação da métrica entropia, e assim por diante. Mais ainda, resultados de acurácia obtidos com o mecanismo de recuperação de casos utilizando sub-bases de casos oriundas do processo de clustering usando funções de similaridade ajustadas podem ser comparados a resultados de acurácia tomados como baseline (armazenados em memória, assim como descrito no passo C anterior).

Figura 3 – Atividades de integração entre CBR e clustering assim como exploradas no estudo de caso desenvolvido nesta dissertação.



4.3 RESULTADOS DOS TESTES REALIZADOS NO ESTUDO DE CASO DO PROJETO SIS-ASTROS DESENVOLVIDO NESTA DISSERTAÇÃO

Na Tabela 1, resultados dos testes realizados no problema de aplicação utilizado neste estudo de caso, assim como descritos nas seções 4.2.3, 4.2.4 e 4.2.5, são apresentados. Na Tabela 2, resultados relacionados a testes feitos em CBR, assim como descritos na Seção 4.2.6, são apresentados. A Tabela 1 apresenta resultados das diferentes métricas de avaliação de grupos de casos, onde estes grupos foram obtidos quando os algoritmos de densidade, hierárquico e particional foram executados. Nestes testes, os esquemas de avaliação de relevância "idêntico" (Tabela 1 – coluna a), "individual" (Tabela 1 – coluna b) e "ajustado" (Tabela 1 – colunas c, d) foram explorados. Em seguida, resultados associados ao esquema de avaliação de relevância "individual" são descritos (Tabela 1 – colunas b1, ..., bn). Ao final, resultados do esquema de avaliação de relevância "ajustado" são descritos (Tabela 1 – colunas c1, c2, d1, d2). No Anexo A, os resultados de clustering parcialmente descritos na Tabela 1 são apresentados de forma completa.

Na Tabela 2, resultados de avaliação de cross-validation em CBR são apresentados. Esses resultados permitem visualizar a acurácia do sistema CBR conforme cada técnica utilizada para estimar a importância relativa de atributos utilizados nas computações de similaridade entre casos. Em contraste, pode-se comparar esses resultados com a utilização de sub-bases de casos na resposta de consultas CBR. Portanto, essa tabela apresenta resultados relativos a a) execução de CBR utilizando todos os atributos com peso = 1.00 (Tabela 2 – coluna a1.1), onde sub-bases de casos não são usadas na resposta de consultas CBR; b) execução de CBR utilizando todos os atributos com peso = 1.00 (Tabela 2 – coluna a1.2), onde as sub-bases formadas em clustering com a utilização de uma função de similaridade onde todos os atributos têm peso = 1.00 são utilizadas; c) execução de CBR utilizando um conjunto de pesos obtidos segundo as técnicas de análise de índices exploradas no ajuste de atributos (Tabela 2 – coluna b1.1) (utilizando o método linear, sem levar em conta valores baseline na formação de pesos), onde sub-bases de casos não são usadas na resposta de consultas CBR; d) execução de CBR utilizando um conjunto de pesos obtidos segundo as técnicas de análise de índices exploradas no ajuste de atributos (Tabela 2 – coluna b1.2) (utilizando o método linear, sem levar em conta o baseline para a formação de pesos), onde as sub-bases de casos formadas em clustering com a utilização de uma função de similaridade ajustada são utilizadas. As demais colunas da Tabela 2 seguem o mesmo padrão de descrição. Porém, elas descrevem valores obtidos quando o

método de normalização logarítmico foi explorado na formação de pesos utilizados nas computações de similaridade entre casos.

Tabela 1 – Tabela dos resultados de avaliação dos grupos formados com a execução dos algoritmos de clustering de densidade, hierárquico e particional, a respeito da base de casos do projeto SIS-ASTROS.

Algoritmo de clustering	Parâmetros de entrada	Métricas de avaliação de grupos	Esquema de avaliação de relevância idêntico	Esquema de avaliação de relevância individual					Esquema de avaliação de relevância ajustado			
			(a) Todos valores de pesos para atributos de caso são iguais a 1.0 (avaliações de baseline)	(b) Valores de pesos para um atributo de caso selecionado = HIGH_VALUE, enquanto todos outros valores de peso para atributos = 1.0					(c) Valores de pesos derivados da *normalização linear* das avaliações de grupos		(d) Valores de pesos derivados da *normalização logarítmica* das avaliações de grupos	
				(b1) Alto valor de peso para o atrib. 1	…	(b10) Alto valor de peso para o atrib. 10	(b11) Alto valor de peso para o atrib. 11	…	(c1) Resultados de baseline *não* são usados	(c2) Resultados de baseline são usados	(d1) Resultados de baseline *não* são usados	(d2) Resultados de baseline são usados
Densidade – 16 grupos gerados	Eps = 3.5 minpts = 15	Pureza	0.37	0.29	…	0.49	0.47	…	0.32	0.42	0.38	0.43
		Entropia	-227.43	-354.62	…	-322.88	-405.45	…	-244.89	-245.66	-243.11	-251.10
		Precisão	0.48	0.46	…	0.75	0.81	…	0.50	0.48	0.51	0.52
Hierárquico - Ponto de corte = 15 grupos	Linkage = Complete	Pureza	0.08	0.08	…	0.33	0.14	…	0.08	0.09	0.13	0.17
		Entropia	-20.38	-25.89	…	-217.03	-43.05	…	-19.43	-20.79	-20.83	-23.74
		Precisão	0.61	0.80	…	0.92	0.91	…	0.59	0.63	0.64	0.70
	Linkage = Average	Pureza	0.15	0.14	…	0.17	0.18	…	0.15	0.17	0.15	0.17
		Entropia	-35.20	-99.14	…	-145.60	-93.89	…	-34.23	-48.71	-34.71	-48.77
		Precisão	0.53	0.62	…	0.63	0.58	…	0.52	0.64	0.52	0.64
	Linkage = Single	Pureza	0.07	0.09	…	0.23	0.21	…	0.08	0.13	0.08	0.13
		Entropia	-33.11	-52.10	…	-180.34	-165.73	…	-32.07	-40.11	-32.02	-40.08
		Precisão	0.52	0.47	…	0.61	0.68	…	0.48	0.54	0.48	0.54
Particional	K = 15	Pureza	0.10	0.10	…	0.24	0.36	…	0.12	0.14	0.17	0.16
		Entropia	-131.10	-131.22	…	-205.09	-221.18	…	-141.90	-145.68	-142.87	-148.69
		Precisão	0.49	0.47	…	0.69	0.88	…	0.47	0.55	0.54	0.58

Tabela 2 – Tabela dos resultados de cross-validation em CBR para os diferentes algoritmos e métricas de avaliação associadas, a respeito da base de casos do projeto SIS-ASTROS.

Algoritmo de clustering	Métricas de avaliação de grupos	(a) Acurácia obtida com o esquema de avaliação de relevância idêntico		(b) Acurácia obtida com o esquema de avaliação de relevância ajustado (pesos derivados do método de normalização linear)				(c) Acurácia obtida com o esquema de avaliação de relevância ajustado (pesos derivados do método de normalização logarítmico)			
				(b1) Avaliações de baseline não são usadas		(b2) Avaliações de baseline são usadas		(c1) Avaliações de baseline não são usadas		(c2) Avaliações de baseline são usadas	
		(a1.1) Sub-bases de casos não são usadas	(a1.2) Sub-bases de casos são usadas	(b1.1) Sub-bases de casos não são usadas	(b1.2) Sub-bases de casos são usadas	(b2.1) Sub-bases de casos não são usadas	(b2.2) Sub-bases de casos são usadas	(c1.1) Sub-bases de casos não são usadas	(c1.2) Sub-bases de casos são usadas	(c2.1) Sub-bases de casos não são usadas	(c2.2) Sub-bases de casos são usadas
Densidade	Pureza	44.50	50.23	62.12	64.11	74.62	72.55	54.76	55.85	65.78	69.42
	Entropia	44.50	50.23	55.25	60.23	57.12	63.88	53.50	59.85	56.35	65.24
	Precisão	44.50	50.23	65.12	71.90	68.62	73.88	69.19	67.10	69.10	73.50
Hierárquico/Single	Pureza	44.50	53.88	53.25	64.09	60.08	73.51	54.21	63.15	67.24	75.41
	Entropia	44.50	53.88	56.18	69.13	58.96	79.63	57.14	67.88	66.98	76.52
	Precisão	44.50	53.88	57.05	61.22	63.40	71.62	56.98	68.85	63.56	77.86
Hierárquico /Average	Pureza	44.50	60.72	57.52	62.00	61.65	69.23	55.23	61.63	65.20	75.30
	Entropia	44.50	60.72	61.66	67.54	64.14	73.42	56.74	62.47	63.12	72.89
	Precisão	44.50	60.72	53.20	70.12	63.23	79.65	59.80	63.54	58.53	70.25
Hierárquico /Complete	Pureza	44.50	69.26	63.11	70.37	67.63	74.36	55.75	64.65	69.25	78.55
	Entropia	44.50	69.26	65.14	68.50	65.26	82.79	59.73	65.19	62.48	83.93
	Precisão	44.50	69.26	58.38	65.40	68.00	79.88	55.88	71.22	63.78	80.50
Particional	Pureza	44.50	51.07	69.85	76.42	71.00	75.42	54.75	55.99	65.62	69.52
	Entropia	44.50	51.07	68.88	72.26	69.12	73.00	53.40	59.75	56.25	65.23
	Precisão	44.50	51.07	66.21	69.44	72.12	74.56	69.08	67.00	69.15	73.54

4.3.1 Resultados de configuração de parâmetros de entrada dos algoritmos de clustering explorados nesta dissertação

Os resultados demonstrados nesta etapa foram obtidos utilizando os seguintes valores de parâmetros de entrada para cada algoritmo de clustering escolhido: no algoritmo de densidade DBScan, eps = 3.5 e minpts = 15; no algoritmo hierárquico DIANA, linkage = [Average, Complete, Single]; e no algoritmo particional K-Means, k = 15. Além disso, a execução destes algoritmos de clustering resultou na seguinte formação de grupos de casos: algoritmo de densidade DBScan = 16 grupos formados; algoritmo hierárquico DIANA = 15 grupos formados e algoritmo particional K-Means = 15 grupos formados.

É importante notar que, assim como definido nesta dissertação, um número bastante similar, ou idêntico, de grupos de casos foi explorado nos diferentes algoritmos de clustering utilizados no processo de análise de índices proposto. Entre outros motivos, essa dissertação assumiu que um número similar de grupos poderia permitir uma melhor comparação entre resultados obtidos por algoritmos de clustering de diferentes naturezas (tais como os três diferentes algoritmos utilizados nesta dissertação). Contudo, maiores análises neste tipo de suposição (onde um número de grupos não seja fixo entre os diferentes algoritmos de clustering explorados) podem ser realizadas a partir de testes, assim como proposto como trabalho futuro dessa dissertação.

4.3.2 Resultados dos testes realizados em clustering segundo o esquema de avaliação de relevância “idêntico”

Neste passo, resultados de pureza, entropia e precisão entre diferentes algoritmos de clustering são apresentados e discutidos. Neste caso, a pureza obtida utilizando o algoritmo de densidade (0.37) é superior aos outros resultados de pureza obtidos com algoritmos hierárquico (Average, Complete e Single) e particional. A entropia utilizando o algoritmo de densidade (-227.43) é superior aos outros resultados de entropia obtidos com hierárquico (Average, Complete e Single) e particional. A precisão obtida utilizando o hierárquico utilizando Complete linkage (0.61) é superior aos outros resultados de precisão obtidos com os algoritmos de densidade, hierárquico (Average e Single) e particional. No estudo de caso realizado, o melhor resultado de pureza foi obtido com o algoritmo de densidade, o melhor resultado de entropia também foi obtido com o algoritmo de densidade, e o melhor resultado de precisão foi obtido com o algoritmo hierárquico utilizando Complete linkage. Portanto, uma concordância

entre melhores resultados utilizando diferentes algoritmos e métrica de avaliação de grupos de casos não foi encontrada.

Ao explorar individualmente resultados de métricas de avaliação de qualidade de grupo de casos, é possível identificar e usar resultados que sejam muito superiores a outros. Por exemplo, o resultado de pureza obtido com o algoritmo de densidade (Tabela 1 – coluna a) é bastante superior aos resultados de pureza obtidos quando outros algoritmos de clustering foram executados e seus grupos de casos foram avaliados. Portanto, é possível concluir que índices obtidos a partir de resultados de pureza associadas a grupos de casos oriundos do algoritmo de densidade poderiam ser explorados em CBR. Além disso, a entropia obtida com o algoritmo de densidade é bastante superior aos resultados de entropia obtidos com os outros algoritmos de clustering, assim sendo uma boa escolha a ser explorada no processo de construção de índices para CBR. No caso de avaliações de precisão, o melhor resultado obtido com o algoritmo hierárquico utilizando Complete linkage é muito similar aos resultados de precisão obtidos com os outros algoritmos de clustering. Sendo assim, escolhas baseadas nestes resultados de precisão não podem ser realizadas no processo de indexação de CBR.

4.3.3 Resultados dos testes realizados em clustering segundo o esquema de avaliação de relevância “individual”

Neste passo, o objetivo é apresentar e discutir resultados de avaliação de grupos de casos formados via clustering onde a função de similaridade utilizada é configurada de acordo com resultados obtidos quando o esquema de avaliação de relevância “individual” é executado. Resultados de avaliação de grupos formados com a execução do esquema de avaliação “individual” podem então ser investigados de diferentes formas: a) análise entre grupos de casos relacionados a diferentes atributos altamente ponderados na função de similaridade; b) análise entre grupos relacionados a execução de diferentes algoritmos de clustering e c) análise de grupos relacionados a atributos altamente ponderados na função de similaridade em relação a análise de grupos baseline relacionados a todos atributos com pesos idênticos e iguais a 1.00.

A) Resultados de pureza, entropia e precisão de grupos de casos obtidos quando o esquema de avaliação “individual” foi executado - análise entre grupos relacionados a diferentes atributos altamente ponderados na função de similaridade

A partir dos resultados obtidos, é possível comparar os resultados de atributos para cada algoritmo e métrica associada (Tabela 1 – colunas b1, ..., bn). Segue exemplos destas comparações, onde os resultados mínimo e máximo são identificados.

No algoritmo de densidade, o atributo 9 apresenta o maior resultado de pureza (0.53) entre os atributos, enquanto o menor resultado logo acima do baseline é encontrado no atributo 7 (0.37) e o menor resultado sem considerar o baseline é encontrado no atributo 5 (0.24). Uma análise de relevância dos demais atributos pode então ser realizada em relação a estes. Por exemplo, os atributos que têm os maiores valores de pureza acima do baseline são os atributos 3, 9 – 11. De acordo com a métrica entropia, o atributo 11 (-405.45) é o mais relevante, enquanto o atributo 4 (-245.19) é o menos relevante logo acima do baseline e o atributo 5 (-195.66) é o menos relevante sem considerar o baseline. Na métrica entropia, os atributos 1, 9 – 13 têm os maiores valores de entropia. De acordo com a métrica precisão, o atributo 9 (0.90) é o mais relevante, enquanto os atributos 5 e 13 são os menos relevantes logo acima do baseline (0.51) e os atributos 3 e 4 (0.42) são os menos relevantes sem considerar baseline. Na métrica precisão, os atributos 2, 9 – 11 têm os maiores valores de precisão.

Em geral, todos os atributos que possuem maior relevância podem ser associados a um peso maior em relação aos demais atributos usados na função de similaridade ajustada empregada em CBR. Análises de resultados de avaliação de qualidade de grupos de casos podem ser realizadas de maneira similar de acordo com os demais algoritmos de clustering e métricas explorados.

Os resultados demonstrados acima do baseline têm relevância no processo de definição de pesos para atributos de casos, onde atributos podem ser definidos como relevantes e irrelevantes a partir da investigação de valores baseline. Portanto, os maiores e menores valores obtidos devem ser identificados para serem explorados no processo de normalização, onde pesos são definidos para serem usados em funções de similaridade ajustadas. Além disso, esses maiores e menores valores encontrados por diferentes métricas e algoritmos de clustering podem estar relacionados a diferentes atributos, podendo dificultar a definição de valores de pesos que sejam consensuais entre diferentes métricas e algoritmos de clustering.

B) Resultados de pureza, entropia e precisão de grupos de casos obtidos quando o esquema de avaliação “individual” foi executado - análise entre grupos relacionados a execução de diferentes algoritmos de clustering

Nesta comparação, pode-se comparar os resultados de atributos entre cada algoritmo, conforme as métricas (Tabela 1 – colunas b1, ..., bn). Segue exemplos disso.

Os maiores resultados de pureza aparecem em: algoritmo de densidade = atributo 9 (0.53); hierárquico utilizando Average linkage = atributo 9 (0.29); hierárquico utilizando Complete linkage = atributo 9 (0.39); hierárquico utilizando Single linkage = atributo 10 (0.23); e particional = atributo 11 (0.36). Já os menores resultados de pureza aparecem em: densidade = atributo 5 (0.24); hierárquico utilizando Average linkage = atributos 6 e 8 (0.08); hierárquico utilizando Complete linkage = atributos 5 e 8 (0.05); hierárquico utilizando Single linkage = atributo 5 (0.04); e particional = atributos 4, 7, 8, 12 e 13 (0.09).

Utilizando o resultado de baseline de cada algoritmo como referência (Tabela 1 - coluna a), é possível identificar quantos atributos podem ser associados a pesos maiores que 1.00 em uma função de similaridade ajustada. Esses resultados também podem indicar uma escala de pesos para esses atributos. Além disso, os resultados obtidos permitem identificar atributos que possivelmente não sejam relevantes para as computações de similaridade entre casos, visto que as avaliações de qualidade associadas a tais atributos estão abaixo do baseline. Por exemplo, resultados de precisão obtidos com a avaliação dos grupos de casos formados com o algoritmo hierárquico utilizando Complete linkage indicam que existem 6 atributos com valores de precisão acima do baseline (baseline de precisão = 0.61): atributo 1 (0.80), atributo 2 (0.69), atributo 9 (0.94), atributo 10 (0.92), atributo 11 (0.91) e atributo 13 (0.62). Na prática, esses atributos que têm valores de precisão acima do baseline são candidatos a serem associados a pesos maiores que 1.00 na função de similaridade utilizada.

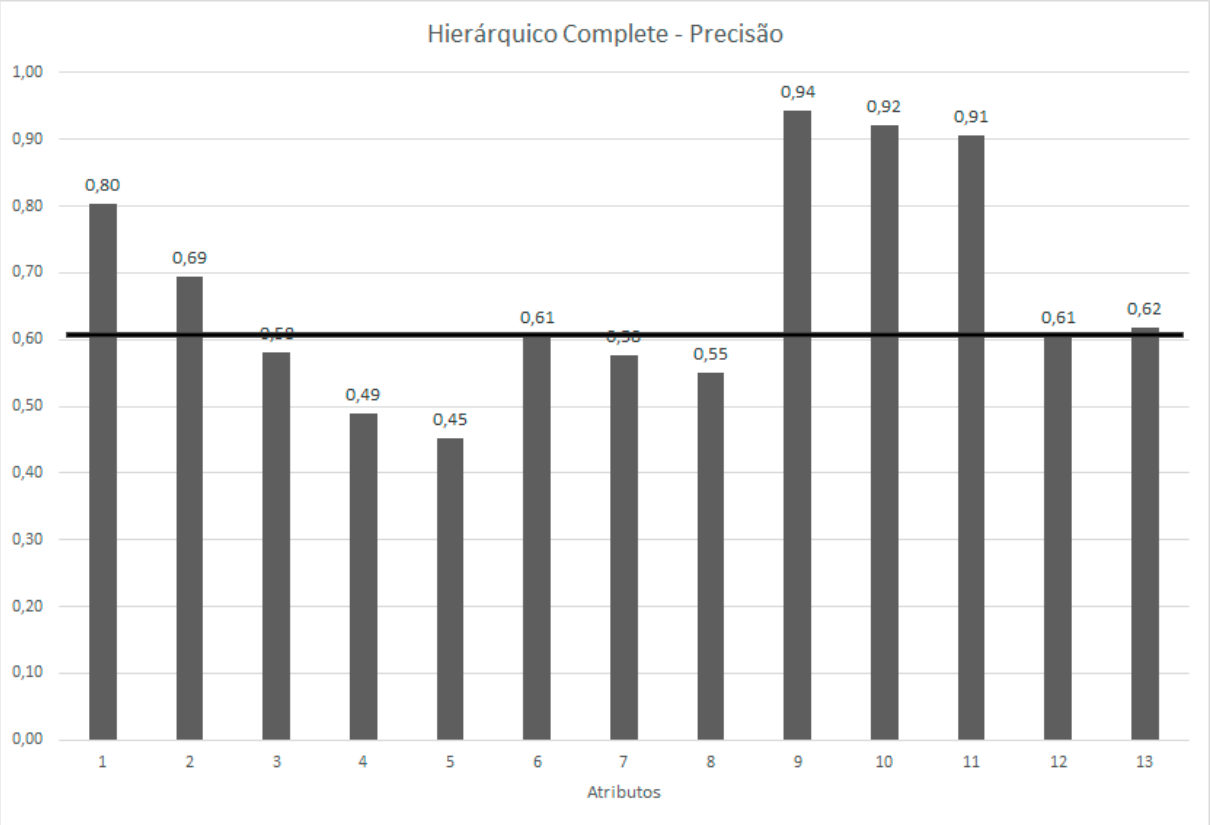
C) Resultados de pureza, entropia e precisão de grupos de casos obtidos quando o esquema de avaliação “individual” foi executado - análise de grupos relacionados a atributos altamente ponderados na função de similaridade em relação a análise de grupos baseline relacionados a todos atributos com pesos idênticos e iguais a 1.00

A partir dos resultados obtidos, valores de avaliação de entropia, pureza e precisão obtidos com pesos = 100.00 em cada atributo selecionado (Tabela 1 – coluna b1, ..., bn) podem ser comparados com valores de avaliação obtidos quando todos os atributos foram associados ao peso = 1.00 na função de similaridade (Tabela 1 – coluna a). Por exemplo, a linha horizontal do gráfico da Figura 2 apresenta um valor baseline de precisão. Nesta figura, resultados de precisão obtidos com a avaliação dos grupos de casos formados com o algoritmo hierárquico utilizando Complete linkage podem ser comparados ao baseline de precisão = 0.61. Sendo assim, os atributos 1 (0.80), 2 (0.69), 9 (0.94), 10 (0.92), 11 (0.91) e 13 (0.62) podem ser tomados como candidatos a receberem um peso maior que 1.00 na função de similaridade

utilizada em CBR. Em contraste, os atributos 3 (0.58), 4 (0.49), 5 (0.45), 6 (0.61), 7 (0.58), 8 (0.55) e 12 (0.61) são candidatos a terem o peso = 1.00, assim como definido nesta dissertação.

Em resumo, a partir dos testes realizados na aplicação explorada no estudo de caso, é possível observar resultados semelhantes nos três diferentes algoritmos de clustering. Tais resultados podem ser utilizados na identificação de atributos que podem ter uma maior relevância (ou peso) nas computações de similaridade entre casos, tanto em CBR quando em clustering. Por exemplo: a) de acordo com o algoritmo de densidade, os atributos 1 – 3, 9 – 13 são os mais relevantes; b) de acordo com o algoritmo hierárquico utilizando Complete linkage, os atributos 1, 3, 9 – 11 são os mais relevantes e c) no algoritmo particional, os atributos 2, 3, 6, 9 – 11, 13 são os mais relevantes. Em geral, tal relevância indica que estes atributos têm uma contribuição mais positiva na formação de grupos de casos mais homogêneos, assim como explorado no estudo de caso desenvolvido nesta dissertação.

Figura 4 – Resultados de avaliação de qualidade de grupos de casos[1].



[1]Resultados da base de casos do projeto SIS-ASTROS obtidos com a métrica precisão na avaliação de grupos formados a partir da execução do algoritmo hierárquico utilizando Complete linkage.

4.3.4 Resultados dos testes realizados em clustering segundo o esquema de avaliação de relevância "ajustado"

Neste passo, primeiramente, a partir dos resultados de avaliação de qualidade de grupos de casos obtidos no passo 4.3.3 anterior, métodos de normalização são usados para a definição de valores de pesos para atributos usados nas computações de similaridade entre casos. Por exemplo, nos resultados de entropia associados aos grupos de casos formados com o algoritmo particional, o primeiro maior valor superior ao baseline é -131.22, sendo associado ao valor de relevância = 1.01 em um método de normalização linear. De forma similar, este valor é associado ao valor de relevância = 1.78 em um método de normalização logarítmico. O maior valor acima do valor baseline é -221.18, sendo associado ao valor máximo de relevância = 10.00. No caso de não considerar o baseline na normalização dos resultados de avaliação de grupos, neste caso, o valor inferior selecionado é -127.92, sendo associado ao valor de

relevância = 1.00, enquanto o maior valor -221.18 permanece o mesmo neste processo de normalização.

Na prática, as estimativas de relevância obtidas a partir do processo de normalização diretamente podem indicar a relevância de cada atributo utilizado nas computações de similaridade, assim como explorado nesta dissertação. Atributos com estimativas de relevância iguais ou muito próximas a 1.00, por exemplo, podem ser considerados como tendo pouca relevância nestas computações de similaridade. Tais atributos, inclusive, podem ser completamente excluídos da função de similaridade usada, apesar das possíveis consequências deste tipo de exclusão não ter sido testado nesta dissertação.

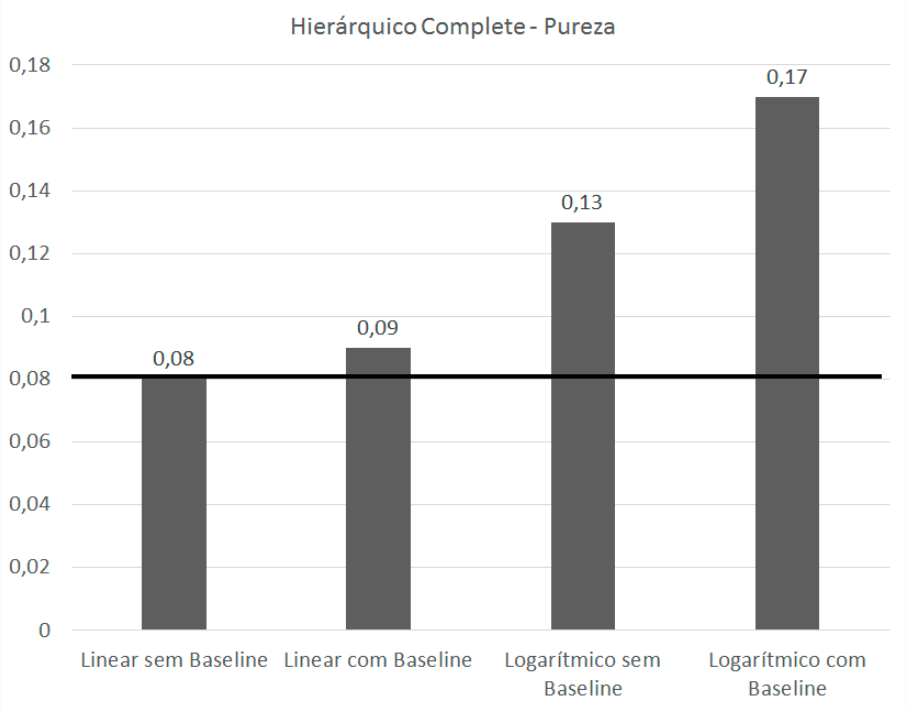
Na Tabela 1, é possível notar que melhorias nos resultados de avaliação de grupos de casos formados foram obtidas. Segue exemplos para cada algoritmo.

a) No algoritmo de densidade, apenas o método de normalização linear (Tabela 1 – coluna c1) apresenta resultados inferiores ao baseline;
b) No algoritmo hierárquico utilizando Complete linkage, apenas o método de normalização linear (Tabela 1 – coluna c1) apresenta resultados inferiores ao baseline;
c) No algoritmo hierárquico utilizando Average linkage, resultados superiores ao baseline são obtidos apenas quando o valor baseline é utilizado no processo de normalização (Tabela 1 – colunas c2, d2);
d) No algoritmo hierárquico utilizando Single linkage, apenas o método de normalização linear (Tabela 1 – coluna c1) apresenta resultados inferiores ao baseline;
e) No algoritmo particional, apenas o método de normalização linear (Tabela 1 – coluna c1) apresenta resultados inferiores ao baseline.

Como resultado da exploração dos esquemas de avaliação de relevância propostos nesta dissertação (esquema de avaliação de relevância “idêntico”, esquema de avaliação de relevância “individual” e esquema de avaliação de relevância “ajustado”), assim como apresentados na Tabela 1, foi possível observar a influência do emprego de estimativas de relevância na configuração de funções de similaridade utilizadas em clustering. Mais ainda, pode-se observar que grupos de casos formados a partir da execução dos algoritmos de clustering onde a função de similaridade utilizada foi ajustada, de acordo com o esquema de avaliação de relevância “ajustado”, podem ser diferentes quando comparados com grupos formados usando uma função de similaridade baseada no esquema de avaliação de relevância “idêntico”.

A influência do uso de pesos em agrupamentos de casos pode ser vista, por exemplo, no gráfico presente na Figura 5. Nessa figura, a linha horizontal representa o resultado de avaliação de grupos de casos utilizando a métrica pureza, quando os casos foram agrupados sem utilizar pesos na função de similaridade. Esta linha horizontal pode ser contrastada com os resultados de pureza obtidos ao utilizar pesos oriundos de normalização linear e logarítmica considerando ou não o baseline como ajuste da função de similaridade. No caso da Figura 5, utilizar pesos formados com o método de normalização logarítmico e considerando o baseline consegue melhores resultados de pureza ao avaliar os grupos criados.

Figura 5 – Resultados de avaliação do esquema "ajustado" [1].



[1]Resultados da base de casos do projeto SIS-ASTROS obtidos com a métrica pureza na avaliação de grupos formados a partir da execução do algoritmo hierárquico utilizando Complete linkage.

Em geral, os resultados de avaliação de grupos de casos formados com a execução de um esquema ajustado que considera o valor baseline no processo de normalização, independente do método de normalização utilizado, seja linear ou logarítmica, tende a apresentar bons resultados na investigação de formação de grupos de casos mais homogêneos, assim como avaliados via diferentes métricas de avaliação de qualidade de grupos de casos exploradas nesta dissertação.

4.3.5 Resultados dos testes realizados em CBR utilizando pesos obtidos com o esquema de avaliação de relevância “idêntico”

A acurácia de consultas CBR é 44.50%, onde o esquema de avaliação de relevância idêntico foi utilizado nas computações de similaridade entre casos. Este resultado tomado como baseline foi obtido sem utilizar sub-bases de casos. Além disso, todos os atributos foram associados ao peso = 1.0 (Tabela 2, coluna a1.1) na função de similaridade utilizada pelo algoritmo de recuperação de casos em CBR.

4.3.6 Resultados dos testes realizados em CBR utilizando pesos obtidos com o esquema de avaliação de relevância “ajustado”

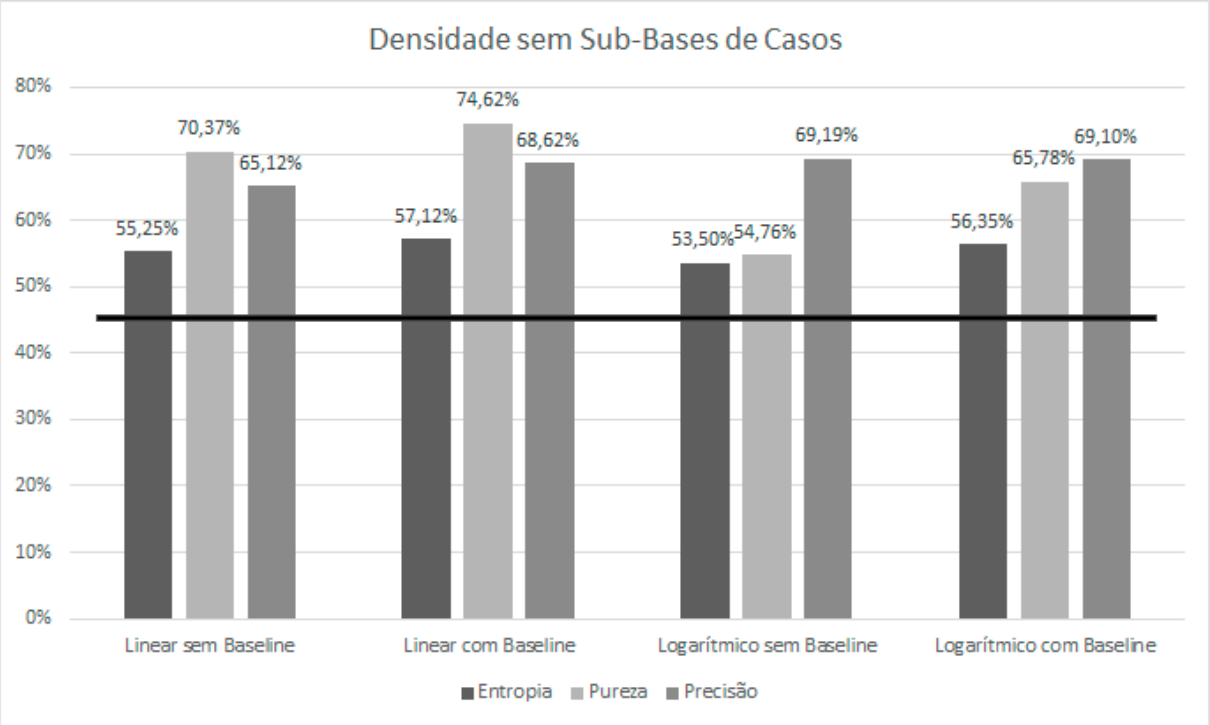
Neste passo, o objetivo é apresentar e discutir resultados de acurácia de consultas CBR onde a função de similaridade utilizada é configurada de acordo com o esquema de avaliação de relevância “ajustado”. Entre outros motivos, a ideia é tentar identificar qual algoritmo de clustering e métrica de avaliação de grupos de casos permite derivar um conjunto de pesos que quando utilizados possam produzir melhores resultados de acurácia nestas consultas CBR.

Por exemplo, a partir de pesos de atributos ajustados na função de similaridade utilizada em CBR, onde tais pesos foram obtidos a partir das execuções do algoritmo de densidade, e grupos de casos resultantes desse algoritmo foram analisados de acordo com a métrica entropia, e os valores resultantes dessas análises de qualidade foram normalizados linearmente sem utilizar o valor baseline neste processo de normalização (Tabela 2 – coluna b1.1), uma acurácia de 55.25% foi obtida. Em contraste, ao usar o valor baseline (Tabela 2 – coluna b2.1) no processo de normalização, uma acurácia de 57.12% foi obtida. Ao usar a normalização logarítmica sem baseline (Tabela 2 – coluna c1.1), uma acurácia de 53.50% foi obtida, enquanto ao usar o baseline (Tabela 2 – coluna c2.1) uma acurácia de 56.35% foi obtida. Similarmente, outras análises como estas podem ser realizadas utilizando os valores apresentados na Tabela 2.

Em geral, analisando os resultados de avaliação de consultas CBR, quando o mecanismo de consultas não utiliza sub-bases de casos, é possível constatar que os melhores resultados são obtidos com pesos formados pelo algoritmo de densidade. Portanto, no gráfico presente na Figura 6 pode-se comparar resultados de acurácia obtidos utilizando pesos formados segundo as métricas entropia, precisão e pureza e normalizados utilizando os métodos linear e

logarítmico considerando o baseline ou não. Além disso, é possível contrastar estes resultados de acurácia obtidos com o baseline que é representado pela linha horizontal do gráfico, onde é notável que ao utilizar pesos neste sistema CBR é possível melhorar a acurácia.

Figura 6 – Resultados de avaliação de consultas CBR do algoritmo de densidade[1].



[1]Resultados da base de casos do projeto SIS-ASTROS obtidos com o mecanismo de consultas ajustado com pesos formados com o algoritmo de densidade e sem utilizar sub-bases de casos.

4.3.7 Resultados dos testes realizados em CBR utilizando pesos obtidos com o esquema de avaliação de relevância “idêntico” e utilizando sub-bases de casos

Neste passo, o objetivo é apresentar e discutir resultados de acurácia de consultas CBR onde o mecanismo de recuperação de casos utiliza sub-bases de casos. Tais sub-bases de casos foram criadas a partir da utilização de um esquema “idêntico” de pesos para os atributos usados nas computações de similaridade entre casos. Assim como descritos na Tabela 3, os resultados de acurácia obtidos com a execução dessas consultas CBR utilizando sub-bases de casos são diferentes entre cada algoritmo de clustering utilizado na execução do esquema de avaliação de relevância “idêntico”. Isso deve-se ao fato de que cada um destes algoritmos de clustering,

densidade, hierárquico e particional, produz diferentes grupos de casos, os quais então são utilizados na obtenção de respostas de consultas CBR.

Tabela 3 – Resultados de acurácia de consultas CBR obtidos com a execução do esquema de avaliação de relevância "idêntico" utilizando sub-bases de casos.

Algoritmo de clustering	(a) Acurácia obtida com o esquema de avaliação de relevância idêntico	
	(a1.1) Sub-bases de casos não são usadas na execução de consultas CBR	(a1.2) Sub-bases de casos são usadas na execução de consultas CBR
Densidade	44.50	50.23
Hierárquico/Single	44.50	53.88
Hierárquico/Average	44.50	60.72
Hierárquico/Complete	44.50	69.26
Particional	44.50	51.07

Na prática, os resultados de acurácia obtidos com a execução de CBR utilizando sub-bases de casos, as quais são formadas com o esquema "idêntico" (Tabela 3 – coluna a1.2), indicam que a utilização de sub-bases de casos por si só consegue melhorar resultados de acurácia nesta aplicação, visto que a não utilização dessas sub-bases de casos está relacionada a uma acurácia de 44.50% (Tabela 3 – coluna a1.1) (nesse caso, o mecanismo de recuperação de casos não utiliza sub-bases de casos e todos os atributos são associados ao peso = 1.00 na função de similaridade). Além disso, ao configurar o mecanismo de recuperação de casos para executar consultas CBR utilizando sub-bases de casos formadas com o algoritmo hierárquico utilizando Complete linkage, é possível obter os melhores resultados de acurácia de consultas CBR (69.26%). Em particular, este valor 69.26% é o melhor resultado obtido (Tabela 3 – coluna a1.2), assim podendo ser utilizado como baseline para a análise de outros resultados de acurácia associados com a execução de consultas CBR.

4.3.8 Resultados dos testes realizados em CBR utilizando pesos obtidos com o esquema de avaliação de relevância "ajustado" e utilizando sub-bases de casos

Neste passo, o objetivo é apresentar e discutir resultados de acurácia de consultas CBR onde o mecanismo de recuperação de casos utiliza sub-bases de casos. Neste caso, estas sub-bases de casos foram formadas pela utilização de um esquema "ajustado" de pesos para os atributos usados nas computações de similaridade entre casos. Os resultados de acurácia

apresentados são baseados na exploração dos métodos de normalização linear (Tabela 2 – coluna b) e logarítmica (Tabela 2 – coluna c), os quais são utilizados na definição de pesos para esses atributos. Ainda, é possível escolher entre não usar o valor baseline (Tabela 2 – colunas b1.2, c1.2) ou usar o baseline (Tabela 2 – colunas b2.2, c2.2) na normalização dos resultados de pureza, entropia e precisão. Isso resulta na definição de diferentes conjuntos de pesos para os atributos utilizados nas computações de similaridade. Mais ainda, estes resultados podem ser comparados com resultados baseline associados a utilização de sub-bases de casos na computação de resultados para consultas CBR assim como apresentada anteriormente (Tabela 3 – coluna a1.2).

A partir dos resultados de acurácia obtidos no problema de aplicação explorado no estudo de caso, pode-se identificar, por exemplo, que ao utilizar sub-bases de casos formadas com a execução do algoritmo hierárquico utilizando Complete linkage, e com a função de similaridade usada contendo atributos associados a pesos oriundos da métrica entropia e do método de normalização linear sem considerar baseline (b1.2), é possível obter uma acurácia de 68.50%. Em contraste, ao usar utilizar pesos oriundos da métrica entropia e do método de normalização linear considerando o baseline (b2.2), é possível obter uma acurácia de 82.79%. Ao usar pesos oriundos da métrica entropia e do método de normalização logarítmico sem considerar baseline (c1.2), é possível obter uma acurácia de 65.19%, enquanto ao usar pesos oriundos da métrica entropia e do método de normalização logarítmico considerando o baseline (c2.2), é possível obter uma acurácia de 83.93%. Ainda, esses resultados de acurácia obtidos podem ser comparados com o melhor resultado baseline de acurácia deste mesmo algoritmo, o qual é 69.26%, assim obtendo duas situações onde o valor de acurácia obtido no baseline é melhorado para 82.79% e 83.93%.

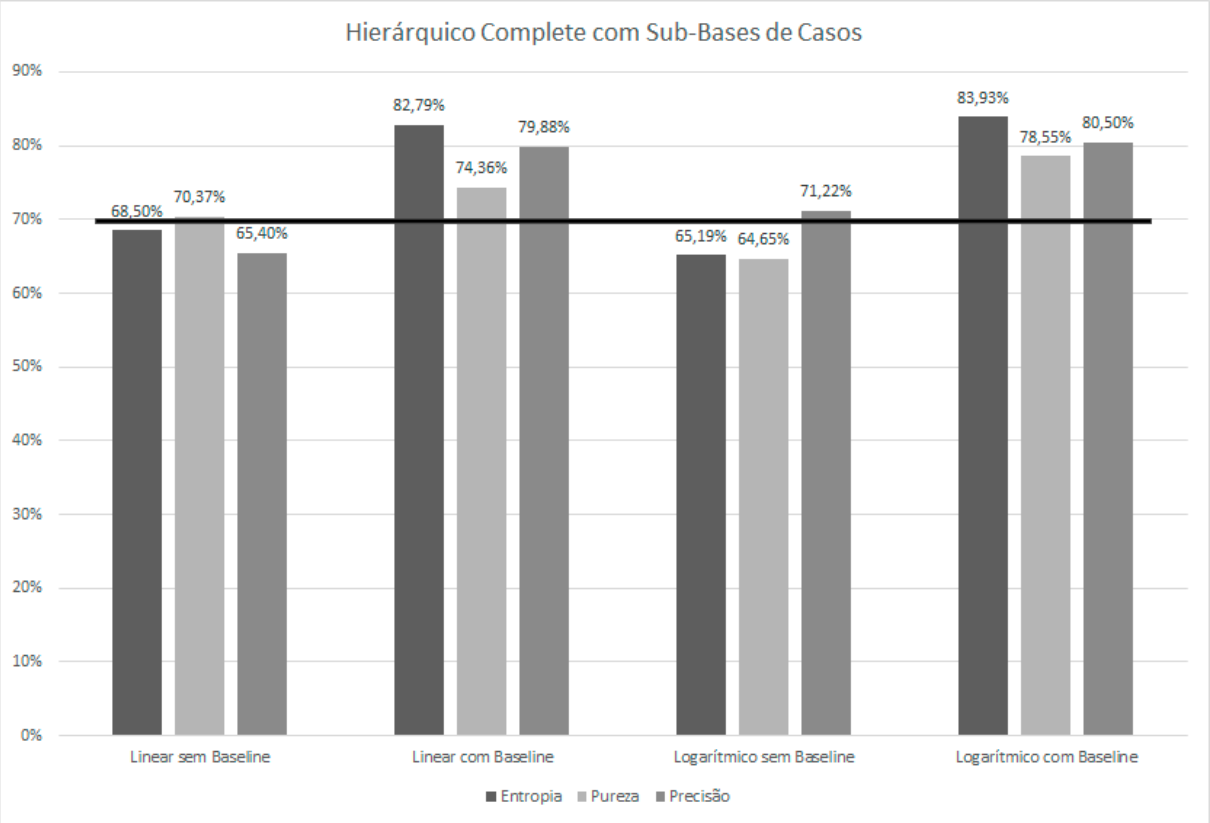
No algoritmo hierárquico utilizando Complete linkage, alguns resultados de acurácia obtidos com a execução de consultas CBR, onde o mecanismo de recuperação de casos utiliza sub-bases de casos oriundas de um esquema de avaliação “ajustado”, apresenta resultados inferiores de acurácia quando comparados ao resultado da execução de consultas CBR utilizando sub-bases de casos oriundas de um esquema de avaliação “idêntico”. Por exemplo, a acurácia obtida com o esquema de avaliação de relevância “ajustado” (pesos derivados do método de normalização linear), quando as avaliações de baseline são usadas no processo de normalização e sub-bases de casos também são usadas (Tabela 2 – coluna b1.2), apresenta resultados inferiores ao baseline (69.26%) em entropia (68.50%) e precisão (65.40%). A acurácia obtida com o esquema de avaliação de relevância “ajustado” (pesos derivados do método de normalização logarítmico), quando as avaliações de baseline não são usadas e sub-

bases de casos são usadas (Tabela 2 – coluna c1.2), apresenta resultados inferiores ao baseline (69.26%) em pureza (64.65%) e entropia (65.19%).

Na prática, os resultados apresentados na Tabela 2 permitem comparar resultados de consultas CBR obtidos quando a) o mecanismo de recuperação de casos não utiliza sub-bases de casos e b) quando o mecanismo de recuperação de casos utiliza sub-bases de casos. Na maioria dos resultados apresentados na Tabela 2, execuções de consultas CBR utilizando sub-bases de casos oriundas de um esquema de avaliação "ajustado" no mecanismo de recuperação de casos apresenta melhores resultados de acurácia quando comparados aos enfoques que não utilizam sub-bases de casos. Além disso, pode-se concluir que ao considerar valores baseline nos métodos de normalização de resultados de avaliação de qualidade de grupos de casos formados em clustering, é possível obter melhores resultados de acurácia. Ao analisar essa tabela, pode-se ainda concluir que utilizar um método de normalização simples, tal como o método linear, resulta em resultados de acurácia melhores que resultados baseline obtidos.

A partir da execução e análise de consultas CBR no estudo de caso explorado nesta dissertação, o melhor resultado de acurácia obtido é de 83.93%. Esse resultado foi alcançado com o mecanismo de consultas CBR utilizando sub-bases de casos. Em particular, tais sub-bases de casos foram oriundas da execução do algoritmo hierárquico utilizando Complete linkage. Este método utiliza uma função de similaridade onde os atributos foram associados a pesos, os quais foram definidos a partir da utilização do método de normalização logarítmica de resultados de entropia de grupos de casos obtidos (e onde um valor baseline foi utilizado no processo de normalização que leva a definição de valores de pesos para atributos). Os resultados de acurácia de consultas CBR oriundos do uso de pesos, quando esses foram formados pelo algoritmo hierárquico utilizando Complete linkage, são então explorados no gráfico da Figura 7. Neste gráfico, é possível contrastar os resultados obtidos ao utilizar pesos com o resultado de acurácia obtido sem o uso de pesos. Dessa forma, claramente nota-se que utilizar pesos eleva consideravelmente a acurácia neste sistema. Portanto, esta configuração do mecanismo de consultas CBR foi a escolhida para apoiar a solução deste problema de aplicação no contexto do projeto SIS-ASTROS. Isso indica que a) as sub-bases de casos encontradas (indexadas a partir de seus centroides) e b) o conjunto de pesos para atributos utilizados nas computações de similaridade estão sendo utilizados como índices no sistema CBR desenvolvido para o problema de aplicação explorado.

Figura 7 – Resultados de avaliação de consultas CBR do algoritmo hierárquico[1].



[1]Resultados da base de casos do projeto SIS-ASTROS obtidos com o mecanismo de consultas ajustado com pesos formados a partir do algoritmo hierárquico e utilizando sub-bases de casos.

Mesmo diante dos resultados de acurácia obtidos, melhorias futuras nestes resultados podem ser direcionadas à consulta de especialistas de domínio no sentido de a) construir funções de similaridade locais específicas do domínio para atributos considerados nas computações de similaridade entre casos e b) coletar um número maior de casos para situações ainda não cobertas pelos casos atualmente disponíveis na base de casos.

4.4 RESULTADOS DOS TESTES REALIZADOS UTILIZANDO BASES DE CASOS DA WEB

O objetivo geral destes testes, os quais complementam os testes desenvolvidos no estudo de caso explorado nesta dissertação, é utilizar diferentes bases de casos disponíveis na Web na avaliação do framework proposto nesta dissertação. Portanto, Tabelas 5 – 10 apresentam resultados de execução de clustering e CBR utilizando bases de casos da Web (LICHMAN, 2013). As bases de casos utilizadas nestas avaliações são: *breast cancer* (**Breast Cancer**

Wisconsin, 1992) com 700 casos representados por 11 atributos; *glass* (**Glass Identification Database**, 1987) com 215 casos representados por 11 atributos; *hepatitis* (**Hepatitis Domain**, 1988) com 155 casos representados por 20 atributos. Sendo essas bases de casos escolhidas por possuírem diferentes números de atributos e pelos dados serem numéricos, semelhante à base de casos do projeto SIS-ASTROS. Nos anexos B, C e D, os resultados de clustering parcialmente descritos nas Tabelas 5, 7 e 9 são apresentados de forma completa.

Tabela 4 – Tabela dos resultados de avaliação dos grupos formados com a execução dos algoritmos de clustering de densidade, hierárquico e particional, utilizando a base de casos de *breast cancer*.

Algoritmo de clustering	Parâmetros de entrada	Métricas de avaliação de grupos	Esquema de avaliação de relevância idêntico	Esquema de avaliação de relevância individual					Esquema de avaliação de relevância ajustado			
			(a) Todos valores de pesos para atributos de caso são iguais a 1.0 (avaliações de baseline)	(b) Valores de pesos para um atributo de caso selecionado = HIGH_VALUE, enquanto todos outros valores de peso para atributos = 1.0					(c) Valores de pesos derivados da *normalização linear* das avaliações de grupos		(d) Valores de pesos derivados da *normalização logarítmica* das avaliações de grupos	
				(b1) Alto valor de peso para o atrib. 1	…	(b4) Alto valor de peso para o atrib. 4	(b11) Alto valor de peso para o atrib. 8	…	(c1) Resultados de baseline *não* são usados	(c2) Resultados de baseline são usados	(d1) Resultados de baseline *não* são usados	(d2) Resultados de baseline são usados
Densidade – 3 grupos gerados	Eps = 3.2 minpts = 2	Pureza	0.19	0.20	…	0.31	0.47	…	0.40	0.42	0.40	0.44
		Entropia	-187.59	-157.54	…	-201.72	-241.39	…	-188.98	-191.55	-186.54	-190.85
		Precisão	0.43	0.52	…	0.70	0.69	…	0.43	0.45	0.42	0.48
Hierárquico - Ponto de corte = 2 grupos	Linkage = Complete	Pureza	0.31	0.40	…	0.62	0.53	…	0.54	0.56	0.55	0.56
		Entropia	-113.26	-125.65	…	-162.10	-138.63	…	-117.52	-117.75	-115.66	-116.63
		Precisão	0.20	0.24	…	0.38	0.31	…	0.22	0.23	0.21	0.24
	Linkage = Average	Pureza	0.31	0.28	…	0.39	0.40	…	0.62	0.64	0.64	0.63
		Entropia	-121.88	-137.63	…	-178.48	-168.30	…	-115.66	-122.36	-119.54	-123.55
		Precisão	0.28	0.37	…	0.44	0.67	…	0.51	0.51	0.50	0.51
	Linkage = Single	Pureza	0.39	0.39	…	0.47	0.49	…	0.61	0.63	0.62	0.63
		Entropia	-98.25	-101.85	…	-130.99	-149.55	…	-99.33	-100.50	-100.12	-110.22
		Precisão	0.40	0.46	…	0.73	0.49	…	0.38	0.40	0.37	0.40
Particional	K = 2	Pureza	0.15	0.26	…	0.45	0.32	…	0.60	0.61	0.58	0.60
		Entropia	-179.65	-182.54	…	-270.98	-239.28	…	-182.69	-183.98	-182.54	-184.66
		Precisão	0.57	0.54	…	0.72	0.74	…	0.66	0.68	0.67	0.61

Tabela 5 – Tabela dos resultados de cross-validation em CBR para os diferentes algoritmos e métricas de avaliação associadas, utilizando a base de casos de *breast cancer*.

Algoritmo de clustering	Métricas de avaliação de grupos	(a) Acurácia obtida com o esquema de avaliação de relevância idêntico		(b) Acurácia obtida com o esquema de avaliação de relevância ajustado (pesos derivados do método de normalização linear)				(c) Acurácia obtida com o esquema de avaliação de relevância ajustado (pesos derivados do método de normalização logarítmico)			
				(b1) Avaliações de baseline não são usadas		(b2) Avaliações de baseline são usadas		(c1) Avaliações de baseline não são usadas		(c2) Avaliações de baseline são usadas	
		(a1.1) Sub-bases de casos não são usadas	(a1.2) Sub-bases de casos são usadas	(b1.1) Sub-bases de casos não são usadas	(b1.2) Sub-bases de casos são usadas	(b2.1) Sub-bases de casos não são usadas	(b2.2) Sub-bases de casos são usadas	(c1.1) Sub-bases de casos não são usadas	(c1.2) Sub-bases de casos são usadas	(c2.1) Sub-bases de casos não são usadas	(c2.2) Sub-bases de casos são usadas
Densidade	Pureza	48.22%	63.52%	59.13%	67.33%	55.22%	67.22%	57.23%	66.54%	55.71%	67.17%
	Entropia	48.22%	63.52%	57.12%	65.99%	61.63%	68.57%	56.57%	65.52%	63.86%	72.58%
	Precisão	48.22%	63.52%	54.97%	64.58%	56.71%	66.87%	55.29%	62.52%	56.20%	67.05%
Hierárquico/Single	Pureza	48.22%	65.55%	53.28%	73.54%	61.77%	78.63%	53.42%	75.68%	62.43%	81.51%
	Entropia	48.22%	65.55%	54.50%	71.56%	55.97%	75.84%	53.76%	74.25%	56.43%	77.20%
	Precisão	48.22%	65.55%	53.54%	73.69%	52.27%	76.58%	54.68%	75.57%	54.70%	76.63%
Hierárquico/Average	Pureza	48.22%	62.66%	53.32%	63.89%	62.03%	71.30%	54.42%	62.90%	61.29%	70.20%
	Entropia	48.22%	62.66%	54.86%	63.54%	61.75%	69.58%	56.00%	64.68%	59.29%	71.52%
	Precisão	48.22%	62.66%	56.50%	66.52%	54.30%	70.30%	56.11%	63.87%	54.71%	70.55%
Hierárquico/Complete	Pureza	48.22%	65.55%	54.14%	66.86%	54.41%	68.85%	53.31%	65.32%	53.43%	69.98%
	Entropia	48.22%	65.55%	54.57%	65.20%	56.53%	70.32%	52.87%	66.32%	59.58%	67.58%
	Precisão	48.22%	65.55%	52.47%	67.01%	52.57%	68.20%	54.55%	66.68%	54.38%	69.52%
Particional	Pureza	48.22%	69.55%	54.78%	71.63%	53.53%	74.54%	55.58%	72.80%	56.42%	76.81%
	Entropia	48.22%	69.55%	52.59%	70.50%	61.21%	78.56%	54.62%	72.52%	57.43%	75.41%
	Precisão	48.22%	69.55%	53.29%	70.56%	61.32%	76.39%	55.90%	75.25%	65.71%	77.58%

Tabela 6 – Tabela dos resultados de avaliação dos grupos formados com a execução dos algoritmos de clustering de densidade, hierárquico e particional, utilizando a base de casos de *glass*.

Algoritmo de clustering	Parâmetros de entrada	Métricas de avaliação de grupos	Esquema de avaliação de relevância idêntico	Esquema de avaliação de relevância individual					Esquema de avaliação de relevância ajustado			
			(a) Todos valores de pesos para atributos de caso são iguais a 1.0 (avaliações de baseline)	(b) Valores de pesos para um atributo de caso selecionado = HIGH_VALUE, enquanto todos outros valores de peso para atributos = 1.0					(c) Valores de pesos derivados da *normalização linear* das avaliações de grupos		(d) Valores de pesos derivados da *normalização logarítmica* das avaliações de grupos	
				(b1) Alto valor de peso para o atrib. 1	…	(b4) Alto valor de peso para o atrib. 4	(b8) Alto valor de peso para o atrib. 8	…	(c1) Resultados de baseline *não* são usados	(c2) Resultados de baseline são usados	(d1) Resultados de baseline *não* são usados	(d2) Resultados de baseline são usados
Densidade – 7 grupos gerados	Eps = 5.5 minpts = 7	Pureza	0.25	0.45	…	0.24	0.39	…	0.25	0.29	0.25	0.32
		Entropia	-131.00	-140.21	…	-110.47	-136.95	…	-132.54	-135.85	-133.20	-133.65
		Precisão	0.37	0.55	…	0.30	0.42	…	0.39	0.40	0.36	0.39
Hierárquico - Ponto de corte = 7 grupos	Linkage = Complete	Pureza	0.42	0.61	…	0.30	0.48	…	0.42	0.42	0.43	0.44
		Entropia	-105.10	-105.10	…	-105.62	-180.80	…	-103.52	-109.63	-105.84	-108.53
		Precisão	0.48	0.48	…	0.38	0.49	…	0.49	0.53	0.48	0.50
	Linkage = Average	Pureza	0.27	0.27	…	0.24	0.35	…	0.28	0.32	0.29	0.31
		Entropia	-98.63	-98.63	…	-95.65	-130.47	…	-99.57	-104.81	-101.85	-100.54
		Precisão	0.44	0.44	…	0.36	0.48	…	0.47	0.46	0.46	0.48
	Linkage = Single	Pureza	0.36	0.36	…	0.32	0.31	…	0.37	0.40	0.38	0.38
		Entropia	-102.56	-102.56	…	-101.68	-127.11	…	-92.85	-96.65	-93.57	-95.54
		Precisão	0.50	0.50	…	0.32	0.50	…	0.50	0.55	0.52	0.53
Particional	K = 7	Pureza	0.15	0.15	…	0.15	0.23	…	0.17	0.18	0.17	0.20
		Entropia	-152.65	-152.65	…	-140.87	-154.63	…	-153.54	-157.63	-155.86	-158.96
		Precisão	0.32	0.32	…	0.21	0.34	…	0.31	0.39	0.35	0.36

Tabela 7 – Tabela dos resultados de cross-validation em CBR para os diferentes algoritmos e métricas de avaliação associadas, utilizando a base de casos de *glass*.

Algoritmo de clustering	Métricas de avaliação de grupos	(a) Acurácia obtida com o esquema de avaliação de relevância idêntico		(b) Acurácia obtida com o esquema de avaliação de relevância ajustado (pesos derivados do método de normalização linear)				(c) Acurácia obtida com o esquema de avaliação de relevância ajustado (pesos derivados do método de normalização logarítmico)			
				(b1) Avaliações de baseline não são usadas		(b2) Avaliações de baseline são usadas		(c1) Avaliações de baseline não são usadas		(c2) Avaliações de baseline são usadas	
		(a1.1) Sub-bases de casos não são usadas	(a1.2) Sub-bases de casos são usadas	(b1.1) Sub-bases de casos não são usadas	(b1.2) Sub-bases de casos são usadas	(b2.1) Sub-bases de casos não são usadas	(b2.2) Sub-bases de casos são usadas	(c1.1) Sub-bases de casos não são usadas	(c1.2) Sub-bases de casos são usadas	(c2.1) Sub-bases de casos não são usadas	(c2.2) Sub-bases de casos são usadas
Densidade	Pureza	49.52%	63.96%	57.24%	71.16%	61.37%	73.23%	67.82%	67.82%	55.15%	71.07%
	Entropia	49.52%	63.96%	57.11%	67.90%	60.32%	74.90%	56.34%	66.23%	54.22%	69.12%
	Precisão	49.52%	63.96%	56.76%	71.74%	65.54%	72.04%	55.80%	66.80%	54.72%	71.26%
Hierárquico/Single	Pureza	49.52%	61.27%	55.65%	65.87%	60.23%	72.98%	56.39%	66.11%	56.85%	68.87%
	Entropia	49.52%	61.27%	56.12%	65.81%	61.98%	67.88%	57.28%	67.08%	55.58%	69.49%
	Precisão	49.52%	61.27%	54.43%	64.46%	57.20%	74.96%	54.58%	65.74%	52.90%	68.31%
Hierárquico/Average	Pureza	49.52%	60.58%	57.43%	68.96%	59.17%	74.99%	56.84%	65.17%	56.05%	68.91%
	Entropia	49.52%	60.58%	56.01%	67.42%	58.04%	72.31%	55.06%	64.31%	55.71%	67.62%
	Precisão	49.52%	60.58%	56.32%	69.11%	61.66%	73.82%	56.97%	66.52%	55.14%	70.82%
Hierárquico/Complete	Pureza	49.52%	64.14%	58.97%	71.68%	61.33%	68.09%	58.31%	69.11%	57.37%	74.45%
	Entropia	49.52%	64.14%	57.46%	70.22%	62.89%	69.11%	57.40%	68.53%	56.63%	72.56%
	Precisão	49.52%	64.14%	55.11%	70.90%	62.09%	71.90%	56.26%	68.14%	54.42%	76.35%
Particional	Pureza	49.52%	67.81%	58.77%	74.58%	60.90%	76.86%	59.02%	71.75%	59.91%	75.21%
	Entropia	49.52%	67.81%	57.22%	73.26%	62.50%	74.09%	58.39%	69.71%	58.16%	74.53%
	Precisão	49.52%	67.81%	56.46%	72.99%	64.99%	73.20%	56.15%	69.02%	60.24%	72.12%

Tabela 8 – Tabela dos resultados de avaliação dos grupos formados com a execução dos algoritmos de clustering de densidade, hierárquico e particional, utilizando a base de casos de *hepatitis*.

Algoritmo de clustering	Parâmetros de entrada	Métricas de avaliação de grupos	Esquema de avaliação de relevância idêntico	Esquema de avaliação de relevância individual					Esquema de avaliação de relevância ajustado			
			(a) Todos valores de pesos para atributos de caso são iguais a 1.0 (avaliações de baseline)	(b) Valores de pesos para um atributo de caso selecionado = HIGH_VALUE, enquanto todos outros valores de peso para atributos = 1.0					(c) Valores de pesos derivados da *normalização linear* das avaliações de grupos		(d) Valores de pesos derivados da *normalização logarítmica* das avaliações de grupos	
				(b1) Alto valor de peso para o atrib. 1	…	(b4) Alto valor de peso para o atrib. 4	(b11) Alto valor de peso para o atrib. 11	…	(c1) Resultados de baseline *não* são usados	(c2) Resultados de baseline são usados	(d1) Resultados de baseline *não* são usados	(d2) Resultados de baseline são usados
Densidade – 2 grupos gerados	Eps = 3.4 minpts = 2	Pureza	0.09	0.06	…	0.20	0.54	…	0.10	0.12	0.10	0.13
		Entropia	-185.56	-170.53	…	-225.62	-267.52	…	-185.56	-190.65	-186.42	-190.12
		Precisão	0.30	0.30	…	0.26	0.46	…	0.32	0.32	0.30	0.32
Hierárquico - Ponto de corte = 2 grupos	Linkage = Complete	Pureza	0.24	0.24	…	0.21	0.52	…	0.27	0.25	0.26	0.37
		Entropia	-220.44	-200.87	…	-260.34	-330.60	…	-215.63	-228.15	-223.52	-228.77
		Precisão	0.42	0.35	…	0.35	0.62	…	0.43	0.44	0.43	0.43
	Linkage = Average	Pureza	0.22	0.25	…	0.25	0.42	…	0.22	0.24	0.23	0.26
		Entropia	-182.91	-221.84	…	-221.12	-265.01	…	-183.45	-182.91	-183.48	-188.56
		Precisão	0.34	0.32	…	0.35	0.58	…	0.35	0.37	0.35	0.39
	Linkage = Single	Pureza	0.18	0.17	…	0.17	0.44	…	0.19	0.19	0.20	0.19
		Entropia	-170.81	-140.30	…	-220.63	-273.44	…	-178.93	-173.80	-173.52	-175.22
		Precisão	0.37	0.38	…	0.40	0.60	…	0.37	0.40	0.37	0.39
Particional	K = 2	Pureza	0.11	0.08	…	0.18	0.32	…	0.10	0.14	0.11	0.14
		Entropia	-190.60	-182.63	…	-203.10	-253.12	…	-192.75	-191.99	-190.55	-191.41
		Precisão	0.24	0.22	…	0.23	0.26	…	0.24	0.27	0.26	0.26

Tabela 9 – Tabela dos resultados de cross-validation em CBR para os diferentes algoritmos e métricas de avaliação associadas, utilizando a base de casos de *hepatitis*.

Algoritmo de clustering	Métricas de avaliação de grupos	(a) Acurácia obtida com o esquema de avaliação de relevância idêntico		(b) Acurácia obtida com o esquema de avaliação de relevância ajustado (pesos derivados do método de normalização linear)				(c) Acurácia obtida com o esquema de avaliação de relevância ajustado (pesos derivados do método de normalização logarítmico)			
				(b1) Avaliações de baseline não são usadas		(b2) Avaliações de baseline são usadas		(c1) Avaliações de baseline não são usadas		(c2) Avaliações de baseline são usadas	
		(a1.1) Sub-bases de casos não são usadas	(a1.2) Sub-bases de casos são usadas	(b1.1) Sub-bases de casos não são usadas	(b1.2) Sub-bases de casos são usadas	(b2.1) Sub-bases de casos não são usadas	(b2.2) Sub-bases de casos são usadas	(c1.1) Sub-bases de casos não são usadas	(c1.2) Sub-bases de casos são usadas	(c2.1) Sub-bases de casos não são usadas	(c2.2) Sub-bases de casos são usadas
Densidade	Pureza	56.23%	73.23%	71.59%	79.32%	77.10%	80.32%	71.15%	78.97%	75.58%	82.98%
	Entropia	56.23%	73.23%	70.23%	78.12%	78.12%	85.55%	72.17%	81.12%	77.69%	85.45%
	Precisão	56.23%	73.23%	72.05%	79.67%	76.15%	80.67%	68.98%	78.64%	74.03%	82.24%
Hierárquico/Single	Pureza	56.23%	70.12%	71.92%	73.50%	78.04%	76.50%	72.30%	76.66%	76.06%	78.28%
	Entropia	56.23%	70.12%	71.68%	74.87%	73.90%	75.87%	69.09%	75.11%	72.09%	77.53%
	Precisão	56.23%	70.12%	72.40%	75.05%	78.88%	76.05%	73.64%	76.90%	77.14%	79.82%
Hierárquico/Average	Pureza	56.23%	72.30%	72.70%	77.32%	77.16%	78.32%	71.28%	76.08%	75.24%	82.42%
	Entropia	56.23%	72.30%	76.67%	79.08%	80.31%	80.08%	75.12%	79.11%	78.19%	84.36%
	Precisão	56.23%	72.30%	77.22%	79.90%	80.90%	83.49%	74.48%	78.33%	80.03%	83.58%
Hierárquico/Complete	Pureza	56.23%	76.29%	70.89%	82.90%	75.43%	83.90%	69.06%	79.09%	74.72%	82.10%
	Entropia	56.23%	76.29%	73.20%	83.14%	81.78%	85.14%	74.04%	80.54%	78.50%	86.01%
	Precisão	56.23%	76.29%	68.71%	80.08%	73.30%	82.08%	69.01%	79.48%	71.63%	83.17%
Particional	Pureza	56.23%	76.11%	69.17%	77.09%	72.12%	81.89%	68.20%	77.97%	72.06%	85.38%
	Entropia	56.23%	76.11%	72.22%	78.40%	71.65%	81.40%	69.09%	79.29%	73.39%	86.60%
	Precisão	56.23%	76.11%	70.05%	76.18%	74.09%	82.18%	70.22%	80.04%	72.36%	86.30%

4.4.1 Discussão dos resultados obtidos com os testes realizados em clustering utilizando as bases de casos de *breast cancer, glass* e *hepatitis*

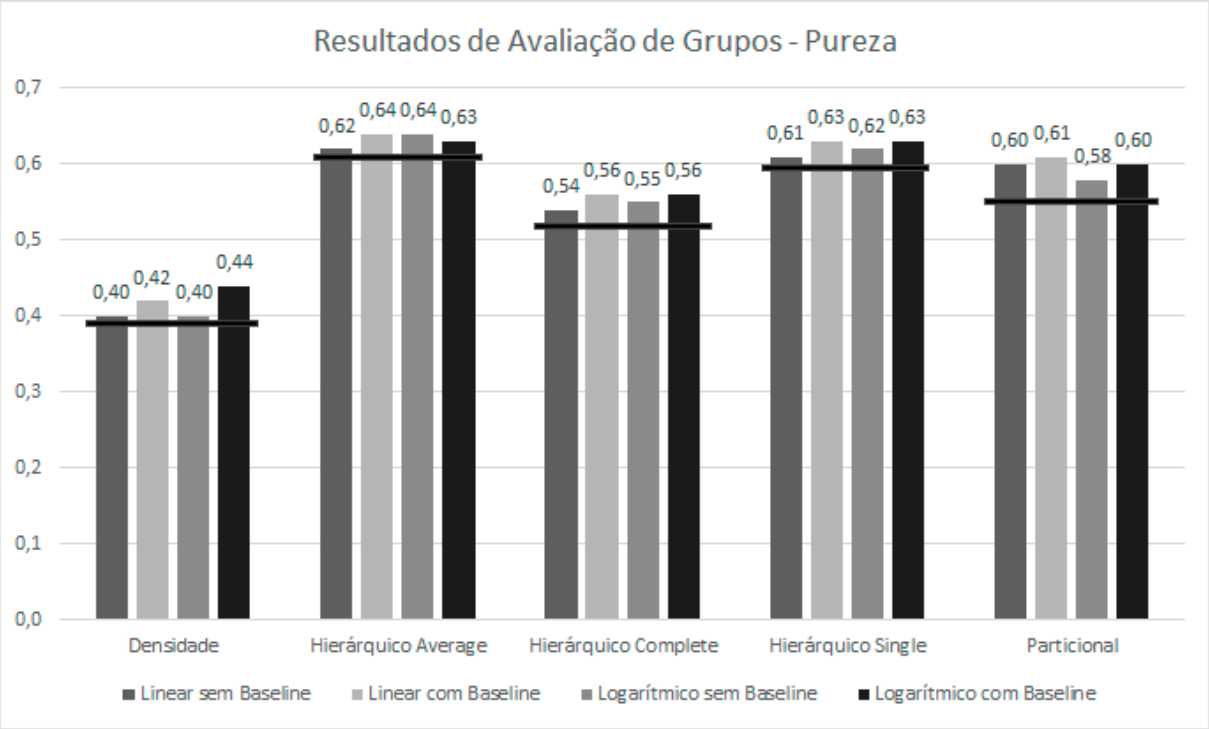
Para ajustar os parâmetros de entrada de cada algoritmo de clustering utilizado, a quantidade de classes (neste caso, rótulos de classes) existentes em cada uma das bases de casos foi considerada na definição do número de grupos a serem formados pelos algoritmos de clustering. Neste sentido, os seguintes parâmetros de entrada dos algoritmos de clustering foram utilizados.

i. *Breast cancer*: a) densidade, eps = 3.4 e minPts = 2; b) hierárquico, nível de corte da árvore = 2; c) particional, k = 2;
ii. *Glass*: a) densidade, eps = 5.5 e minPts = 7; b) hierárquico, nível de corte da árvore = 7; c) particional, k = 7;
iii. *Hepatitis*: a) densidade, eps = 3.4 e minPts = 2; b) hierárquico, nível de corte da árvore = 2; c) particional, k = 2;

Ao analisar resultados da execução do esquema de avaliação de relevância "idêntico" (Tabelas 5, 7, 9 – coluna a), não são encontrados resultados de avaliação de grupos de casos semelhantes entre execuções referentes à cada base de casos. No entanto, estes valores podem ser utilizados como baseline na análise de resultados de avaliação de relevância "individual" associados a cada atributo. Ao analisar valores baseline, é possível notar alguns padrões nos resultados. Por exemplo, na base de casos de *breast cancer*, os atributos 2 – 4, 8 – 9 aparecem como relevantes em todos os testes realizados (neste caso, valores de avaliação para cada um desses atributos são maiores que valores tomados como baseline). Além disso, ao executar o esquema de avaliação de relevância "ajustado" (Tabelas 5, 7, 9 – colunas c, d), foi possível obter melhores resultados em relação ao baseline, como pode ser notado no gráfico da Figura 8.

No gráfico da Figura 8, resultados de avaliação de grupos de casos formados em agrupamentos contendo a função de similaridade ajustada com pesos oriundos da métrica pureza podem ser analisados entre os diferentes algoritmos de clustering. Além disso, estes resultados podem ser comparados em relação aos seus respectivos valores baseline, representado individualmente por uma linha horizontal referente à cada algoritmo. Em resumo, esses resultados permitem afirmar que utilizar pesos na função de similaridade de clustering permite formar grupos mais homogêneos.

Figura 8 – Resultados de avaliação de grupos de casos da base de casos de *breast cancer*[1].



[1]Resultados da base de casos do *breast cancer* obtidos segundo a execução do esquema de avaliação de relevância "ajustado" utilizando pesos formados com a métrica pureza para cada algoritmo de clustering.

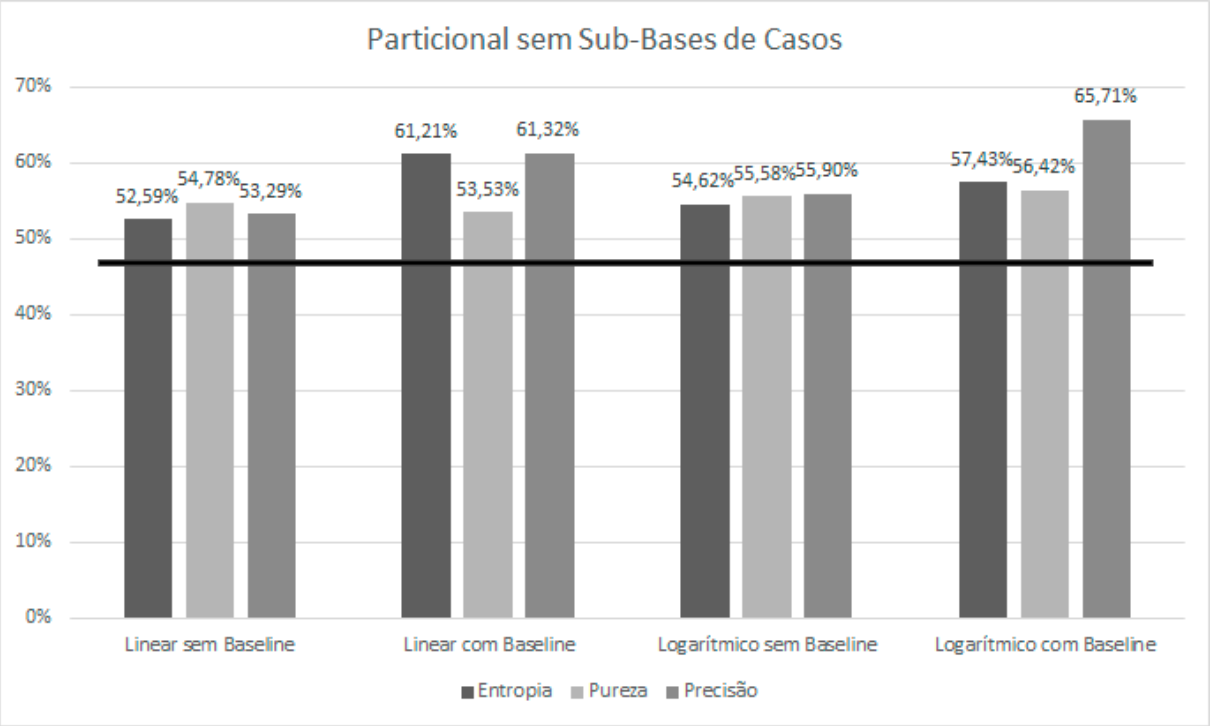
4.4.1 Discussão dos resultados obtidos com os testes realizados em CBR utilizando as bases de casos de *breast cancer*, *glass* e *hepatitis*

Nos resultados de consultas CBR, o objetivo era melhorar os valores baselines de acurácia (Tabelas 6, 8 e 10 – coluna a1.1). Dessa forma, os valores baselines a serem melhorados para cada base de casos são: a) *breast cancer* = 48.22%; b) *glass* = 49.52%; *hepatitis* = 56.23%. Em todos os testes realizados com estas bases de casos, todos os conjuntos de pesos oriundos da execução do esquema de individual de clustering permitiram obter melhores valores de acurácia do que os valores baseline.

Além disso, quando o mecanismo de consulta de casos é configurado para não utilizar sub-bases de casos, cada base de casos nestes problemas de aplicação apresentou um melhor resultado de acurácia obtido com um determinado algoritmo de clustering e um determinado método de normalização, como: a) *breast cancer* = 65.71% (particional e método de normalização logarítmico considerando o baseline); b) *glass* = 65.54% (densidade e método de normalização linear considerando o baseline); *hepatitis* 81.78% (hierárquico e Complete linkage e método de normalização linear considerando o baseline). Um exemplo destes

resultados pode ser visto no gráfico da Figura 9, onde é apresentado o melhor resultado de acurácia sem sub-bases de casos obtido com a base de casos de *breast cancer*.

Figura 9 – Resultados de avaliação de consultas CBR do algoritmo particional sem sub-bases de casos[1].

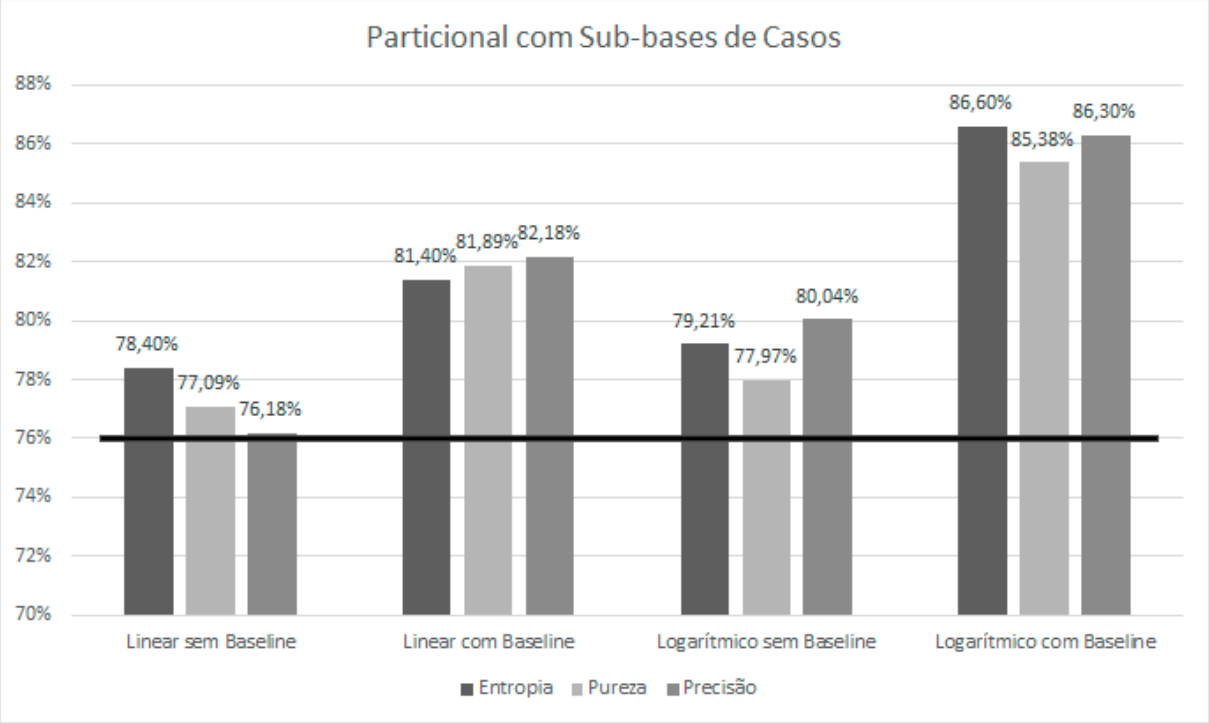


[1]Resultados da base de casos de *breast cancer* obtidos com o mecanismo de consultas de casos similares ajustado com pesos formados a partir do algoritmo particional e sem utilizar sub-bases de casos.

Ainda, é possível melhorar estes resultados de acurácia obtidos quando o mecanismo de consultas de casos similares não utiliza sub-bases de casos. Além disso, quando pesos não são utilizados como ajustes da função de similaridade de CBR e sub-bases de casos são utilizadas no mecanismo de consultas, valores baseline são obtidos (Tabelas 6, 8 e 10 – coluna a1.2) para cada base de casos investigada. Estes valores baselines podem ser utilizados como valores base que servem como comparativos a outros resultados de consultas CBR, nesse caso, quando pesos são empregados na função de similaridade de CBR. Portanto, em seguida, pesos são empregados na função de similaridade com o objetivo de superar estes valores baseline. Os resultados demonstram que é possível melhorar a acurácia de consultas CBR quando o mecanismo de consultas utiliza sub-bases de casos e utiliza pesos como ajustes da função de similaridade. Nesse cenário, os melhores resultados obtidos são: a) *breast cancer* = 81.51% (hierárquico com Single linkage e método de normalização logarítmico considerando o

baseline); b) *glass* = 76.35% (hierárquico com Complete linkage e método de normalização logarítmico considerando o baseline); *hepatitis* 86.60% (particional e método de normalização logarítmico considerando o baseline). Um exemplo desses resultados pode ser visto no gráfico da Figura 10, onde é apresentado o resultado de acurácia com sub-bases de casos obtido com a base de casos de *hepatitis.*

Figura 10 – Resultados de avaliação de consultas CBR do algoritmo particional com sub-bases de casos[1].



[1]Resultados da base de casos de hepatitis obtidos com o mecanismo de consultas ajustado para utilizar sub-bases de casos formadas com o algoritmo particional

Os resultados obtidos nas execuções do nosso framework usando as bases de casos de *breast cancer*, *glass* e *hepatitis* podem ser comparados com resultados oriundos de outras avaliações envolvendo estas três bases de casos (COPERNICUS, 2017). Na base de casos de *breast cancer*, o framework proposto nesta dissertação não conseguiu superar os resultados da literatura (BENNETT; BLUE, 1998; DHIVYAPRIYA; SIVAKUMAR, 2017; FERNÁNDEZ-DELGADO et al., 2014; ŠTER; DOBNIKAR, 1996; YOUH; RUMBE, 2010). Na base de casos de *glass*, o framework desta dissertação conseguiu melhores resultados de acurácia em relação aos resultados apresentados na literatura. Neste caso, o melhor resultado encontrado na literatura é de 75.2% (DOMENICONI; PENG; GUNOPULOS, 2000; FERNÁNDEZ-

DELGADO et al., 2014; LIU et al., 2012), enquanto nosso melhor resultado encontrado de acurácia para essa base de casos é de 76.35%. Na base de casos de *hepatitis*, o framework proposto não conseguiu superar resultados de acurácia encontrados na literatura (ALSHAMRANI; OSMAN, 2017; FERNÁNDEZ-DELGADO et al., 2014; ŠTER; DOBNIKAR, 1996; WEISS; KAPOULEAS, 1990). No entanto, os resultados apresentados estão na média dos resultados disponíveis na literatura, o que também é relevante. Isso justifica-se pelo fato de que os grupos de casos foram organizados de forma a ficarem mais homogêneos. Além disso, os casos organizados em grupos podem ser utilizados em diversas investigações para o usuário conforme sua necessidade. Em geral, a acurácia é um aspecto importante, mas a construção de estruturas de índices e a investigação dos grupos de casos são resultados que complementam em muito os resultados de acurácia apresentados.

4.4 CONSIDERAÇÕES DO CAPÍTULO

Neste capítulo, a construção e análise de funções de similaridade ajustadas e de sub-bases de casos encontradas durante o desenvolvimento desse processo foram investigadas no contexto da indexação de sistemas CBR. Assim como proposto nesta dissertação, esquemas de avaliação de relevância “idêntico”, “individual” e “ajustado” foram executados via algoritmos de clustering visando a definição de estimativas de relevância (descritos como valores de pesos) para atributos utilizados em computações de similaridade entre casos.

De acordo com resultados obtidos no estudo de caso explorado nesta dissertação, a utilização de pesos associados aos atributos utilizados nas computações de similaridade entre casos permitiu obter resultados de acurácia para consultas CBR superiores a resultados tomados como baseline. Quando o mecanismo de consultas CBR não utilizou sub-bases de casos, mas utilizou pesos formados a partir da execução e análise dos algoritmos de densidade, hierárquico e particional, os melhores resultados de acurácia foram obtidos com a utilização da técnica de normalização linear, onde valores tomados como baseline foram utilizados na seleção de um subconjunto de estimativas de qualidade de grupos de casos a serem normalizadas. Neste caso, na aplicação do estudo de caso, o algoritmo de densidade e a métrica pureza apresentaram o melhor resultado de acurácia = 74.62%. Contudo, resultados ainda melhores foram obtidos quando o mecanismo de consultas CBR utilizou sub-bases de casos oriundas da execução de um esquema de avaliação de relevância “ajustado”. A partir de execução do processo de indexação proposto nesta dissertação, o melhor resultado de acurácia alcançado na aplicação CBR explorado no estudo de caso é de 83.93%, alcançado com o mecanismo de consultas CBR

utilizando sub-bases de casos oriundas do algoritmo hierárquico com Complete linkage e com a função de similaridade ajustada com resultados provenientes da normalização (método logarítmico considerando o baseline) de valores obtidos quando a métrica entropia foi utilizada na análise e grupos de casos deste algoritmo de clustering. Além disso, a execução do framework foi repetida utilizando bases de casos disponíveis na Web com o objetivo de comprovar a eficácia de tal técnica. Os resultados demonstraram que este framework realmente consegue criar estruturas de índices capazes de apoiar na construção de funções de similaridade que suportem melhores resultados de acurácia em consultas CBR. Em geral, estes resultados indicam quando o processo de indexação proposto nesta dissertação é explorado, a acurácia de sistemas CBR pode ser melhorada significativamente quando comparada com a acurácia de sistemas CBR que não utilizam estas estruturas de índice.

5 UMA COMPARAÇÃO DA ABORDAGEM DE INDEXAÇÃO PARA CBR PROPOSTA NESTA DISSERTAÇÃO COM TRABALHOS RELACIONADOS

No trabalho de WETTSCHERECK et al. (1997), diferentes enfoques de indexação de bases de casos para melhorar o desempenho de consultas CBR são analisadas. Para isso, os seguintes enfoques são descritos: *feedback de usuários* no ajuste de funções de similaridade; *desempenho de classificações de bases de casos* no que diz respeito ao *tamanho dessas bases; organização de atributos* para substituição de valores anômalos de atributos; *utilização e impacto de funções de similaridade locais e globais em computações de similaridade* para verificar a homogeneidade de grupos criados e *conhecimento relacionado a participação de especialistas na configuração de funções de similaridade*. No entanto, nesse trabalho, não é apresentado um processo que apoia a escolha de pesos para atributos de casos usando clustering. Portanto, de forma a complementar a ideia desse trabalho, esta dissertação propõe um framework de execução de algoritmos de clustering e avaliação de grupos resultantes dessas execuções para a criação de estruturas de índices onde a ideia é descobrir a relevância de atributos de casos usados em computações de similaridade.

No trabalho de ARSHADI; JURISICA (2005), a técnica baseada na mistura de especialistas é relevante para as propostas apresentadas nesta dissertação principalmente quando resultados de classificação de casos de acordo com cada especialista são combinados. De forma a complementar a ideia de mistura de especialistas, esta dissertação apresenta uma proposta que apoia a descoberta de conjuntos de pesos para atributos, os quais podem ser ajustados para refletir o conhecimento de diferentes especialistas envolvidos no processo de

tomada de decisão. Além disso, grupos de casos formados para apoiar cada um desses especialistas poderiam ser mais homogêneos se fossem baseados em funções de similaridade ajustadas, assim como discutido nesta dissertação.

No trabalho de HONG; LIOU (2008), clustering é utilizado para criar grupos de casos para indexar as informações contidas no corpo desses casos. Esses índices criados são usados para corrigir informações anômalas de casos vizinhos existentes nos grupos gerados. Nesse trabalho, no entanto, clustering não é executado utilizando uma função de similaridade ajustada onde a geração de grupos mais homogêneos poderia ser explorada. De forma a complementar a ideia desse trabalho, esta dissertação demonstra como criar uma função de similaridade ajustada para a criação de grupos mais homogêneos. Portanto, os grupos utilizados na indexação dos casos, nesse trabalho, poderiam ser mais homogêneos assim melhorando o desempenho da aplicação utilizando um paradigma como o proposto nesta dissertação. Além disso, esta dissertação também utiliza diferentes algoritmos de clustering para avaliar se o processo de indexação consegue gerar grupos homogêneos utilizando algoritmos de clustering de diferentes naturezas.

No trabalho de YANG; WU (2000), sub-bases de casos são criadas a partir dos grupos formados em clustering, onde centroides desses grupos são usados como índices dessas sub-bases de casos para serem usados nos mecanismos de consultas de CBR. Em VERNET; GOLOBARDES (2003), uma técnica similar a criação de sub-bases de casos é utilizada. Nesse trabalho, grupos criados em clustering são usados na criação de esferas que representam sub-bases de casos. Os casos contidos nessas esferas são novamente agrupados, com o objetivo de tornar essas esferas mais homogêneas. De forma similar, no trabalho de MÜLLER; BERGMANN (2014), a criação de sub-bases de casos é explorada com o algoritmo hierárquico onde um filtro que define os grupos que podem ser pesquisados pelo mecanismo de consultas de casos de CBR é explorado. Para complementar essas ideias, esta dissertação propõe os esquemas de avaliação de relevância (“idêntico” e ”ajustado”), onde sub-bases de casos formadas podem ser mais homogêneas com o uso de funções de similaridade ajustadas em clustering. Além disso, esta dissertação propõe o uso combinado de sub-bases de casos geradas a partir de diferentes algoritmos de clustering e discute resultados de acurácia obtidos.

Na Tabela 5, os trabalhos relacionados são comparados com as propostas apresentadas nesta dissertação.

Tabela 10 – Comparativo de trabalhos relacionados com esta dissertação.

	(ARSHADI; JURISICA, 2005)	(HONG; LIOU, 2008)	(WETTSCHERECK et al., 1997)	(YANG; WU, 2000)	(VERNET; GOLOBARDES, 2003)	(MÜLLER; BERGMANN, 2014)	(LUCCA, 2017)
Qual é o propósito de CBR?	CBR é utilizado para descobrir classes de casos a partir de uma base de casos específica sobre biologia molecular, visando melhor desempenho na etapa recuperação de casos. Estudo de caso desenvolvido em um problema de biologia.	CBR é integrado com clustering para indexar as informações contidas no corpo dos casos, para corrigir informações errôneas contidas nesses casos.	Investigar formas de indexação de bases de casos para melhorar o desempenho de consultas CBR, onde 5 diferentes enfoques de indexação são detalhados.	Investigar como organizar grandes bases de casos para obter melhor desempenho e acurácia em consultas CBR.	Indexação de bases de casos, buscando reduzir o tempo computacional de consultas e evitar outliers em grupos de casos.	Melhorar a fase de recuperação de casos de CBR, utilizando o paradigma POCBR (raciocínio baseado em casos orientado a processos).	Explorar algoritmos de clustering e métricas de avaliação de grupos na construção de funções de similaridade ajustadas e na definição de sub-bases de casos. Estudo de caso desenvolvido em um problema de simulação.
Qual é o propósito de clustering?	Melhorar a acurácia de predição de novos casos. Descobrir quais são os atributos que possuem relacionamento significativo com classes definidas no domínio da aplicação. Algoritmos usados: K-Means e spectral clustering.	Atributos contidos em um mesmo grupo são similares, então o objetivo é mitigar valores anômalos de atributos e melhorar o desempenho de consultas CBR com esta organização de informações em casos.	Este trabalho não utiliza especificamente clustering, mas compara diferentes enfoques de aprendizado de máquina automático para indexação de bases de casos.	Criação de sub-bases de casos a partir de grupos formados em clustering, para serem utilizadas em consultas CBR. Algoritmo usado: CBSCAN baseado no GDBScan.	Usar clustering para indexar bases de casos em sub-bases de casos, para melhorar o desempenho de consultas CBR. A base de casos é então organizada em esferas a partir dos grupos formados em clustering. Algoritmos usados: Mean Sphere e Mean K-Means.	Melhorar desempenho de consultas, possibilitando a recuperação de casos similares em uma estrutura de árvore. Os grupos são definidos como sub-bases de casos nesta estrutura de árvore. Algoritmo usado: HBPAM.	Descoberta de atributos relevantes a partir de diferentes algoritmos de clustering e criação e utilização de sub-bases de casos a partir de grupos de casos formados. Algoritmos usados: densidade (DBScan), hierárquico (DIANA) e particional (K-Means).
Como a indexação é colocada em prática?	Mistura de especialistas, onde sub-bases de casos de cada especialista são utilizadas. Cada especialista classifica casos e os resultados finais são obtidos a partir da maioria de votos. Consultas são então respondidas combinando os resultados de cada especialista.	Indexação dos valores de atributos mais comuns dentro dos grupos formados, definindo estratégias de substituição de valores anômalos entre os casos vizinhos.	Cinco formas de indexação investigadas: feedback de usuários para seleção de atributos relevantes; eliminação de atributos irrelevantes de acordo com grupos formados; ponderação de atributos em funções de similaridade; criação de funções de similaridade locais e globais; e participação de especialistas.	Utilização de centroides como índices de sub-bases de casos, onde usuários participam da escolha de qual é o melhor grupo de casos a ser utilizado em uma consulta CBR dada.	Cada caso é distribuído em esferas (construídas a partir de grupos formados em clustering) que possuem casos semelhantes. As esferas são reagrupadas, para melhorar o desempenho dos índices formados. O mecanismo de consultas de casos utiliza essas esferas na recuperação de casos similares.	Grupos formados com um algoritmo hierárquico são indexados para a investigação de casos similares. Para isto, um filtro com parâmetros que determinam os níveis (grupos) da árvore que podem ser pesquisados é investigado.	Uso de métodos de normalização de resultados de avaliações de grupos de casos para a definição de pesos para atributos usados em computações de similaridade. Criação de sub-bases de casos, onde centroides são utilizados como índices na resposta de consultas CBR.

6 CONCLUSÃO

Sistemas CBR resolvem novos problemas utilizando soluções de casos passados. Para isso, funções de similaridade são usadas na análise da relevância de casos passados no apoio a solução de um dado problema descrito como consulta em sistemas CBR. Adversidades nesta etapa de recuperação de casos de sistemas CBR existem e podem impactar negativamente na acurácia de consultas CBR. Para atacar esses problemas, técnicas de IA podem ser usadas na criação de estruturas de índices visando refletir a relevância de atributos de casos na construção de melhores mecanismos de consultas. Dentre diferentes índices explorados, como a criação de sub-bases de casos para a organização de bases de casos, esta dissertação descreve como explorar a criação e utilização de funções de similaridade ajustadas.

Nesta dissertação, uma sequência de atividades envolvendo a execução e avaliação de algoritmos de clustering para a criação de estimativas de relevância para atributos utilizados na representação de casos foi explorada. Nestas atividades, primeiramente, clustering foi explorado como um processo tradicional de descoberta de conhecimento. No contexto de indexação em sistemas CBR, o objetivo era determinar valores baseline para comparar com resultados de indexação obtidos em outras atividades de análise de índices propostas na dissertação. Em seguida, clustering foi explorado de forma não tradicional, visto que funções de similaridade contendo atributos ponderados de diferentes formas foram usadas. Com isso, algoritmos de clustering foram executados e avaliados visando obter feedback de cada um dos atributos usados nas computações de similaridade entre casos. Essa exploração foi realizada com a utilização de funções de similaridade ajustadas onde cada um dos atributos foi altamente ponderado. Ao fazer isso, a principal suposição foi de que o impacto de um atributo sendo altamente ponderado na função de similaridade usada poderia ser analisado nos grupos de casos resultantes da utilização dessa função ajustada nos algoritmos de clustering. Por último, clustering foi novamente explorado usando o feedback obtido a respeito de possíveis pesos para atributos na configuração da função de similaridade.

A escolha das métricas de avaliação de grupos a serem utilizadas não foi uma tarefa fácil, pois não existe uma forma genérica de avaliação de grupos de casos. Neste caso, métricas externas (pureza, entropia e precisão) foram escolhidas, pois casos da base de casos contêm soluções associadas, as quais foram utilizadas na computação dessas métricas. Com isso, grupos obtidos poderiam ser avaliados em relação a estes “rótulos” de casos. Um outro problema encontrado foi descobrir um valor de peso para ajustar a relevância de atributos na função de similaridade. Na prática, tal valor de peso deveria ter uma influência nos grupos de casos

resultantes dos algoritmos de clustering. Por fim, um dos principais problemas encontrados nesta pesquisa foi conseguir gerar valores de pesos a partir das avaliações de qualidade de grupos obtidas. Para isso, um processo de normalização dos valores de avaliação de qualidade obtidos foi explorado utilizando diferentes algoritmos de clustering e diferentes métricas de avaliação. A ideia envolvia verificar se o processo de identificação de pesos funcionava a partir da utilização de técnicas de diferentes naturezas. Em muitos sentidos, uma solução para o problema de identificação de pesos para atributos utilizados em computações de similaridade entre casos foi encontrada quando valores baseline foram utilizados na definição de pesos. Algoritmos de clustering variados também foram explorados na criação de sub-bases de casos, visto que o objetivo não era só descobrir uma função de similaridade ajustada que proporcionasse uma maior acurácia de consultas CBR. A ideia explorada foi também utilizar clustering para identificar grupos de casos que pudessem ser usados como índices em sistemas CBR, assim como explorado no estudo de caso realizado nesta dissertação.

O processo de análise de índices proposto nesta dissertação foi avaliado em um estudo de caso inserido no contexto de um projeto de um sistema de simulação para o Exército Brasileiro – SIS-ASTROS. Para isso, diferentes implementações foram realizadas para a definição dos algoritmos de clustering, das métricas de avaliação de qualidade de grupos e do processo de avaliação do sistema CBR. O processo de melhorar a acurácia em consultas CBR foi explorado utilizando funções de similaridade construídas. Para isso, diferentes execuções e avaliações de consultas CBR na aplicação de simulação tratada no estudo de caso foram realizadas. Primeiramente, a acurácia de consultas CBR foi avaliada sem usar a função de similaridade construída, usando pesos iguais e idênticos em todos os atributos de casos. Em seguida, a acurácia foi avaliada com o mecanismo de recuperação de casos usando a função de similaridade construída. Por último, a acurácia foi avaliada com o mecanismo de recuperação de casos usando ou não as sub-bases de casos construídas como resultado do processo de clustering proposto na dissertação. O objetivo dessas avaliações foi obter resultados de acurácia para permitir escolher as melhores estruturas de índices a serem usadas na construção do sistema de simulação descrito no projeto SIS-ASTROS. Além disso, as execuções realizadas neste estudo de caso foram repetidas em bases de casos disponíveis na Web, onde resultados comprovaram a eficácia do framework em criar índices capazes de apoiar na construção de funções de similaridade ajustadas para suportar melhores resultados de acurácia em consultas CBR.

Em resumo, a principal contribuição desta dissertação foi propor um novo processo envolvendo a execução e avaliação de algoritmos de clustering visando a investigação e solução

de problemas de indexação em sistemas CBR. Em particular, o framework proposto demonstrou como usar resultados de avaliação de qualidade de grupos de casos na definição de pesos para atributos utilizados em computações de similaridade entre casos. Além disso, um estudo de caso real foi realizado, onde todo o processo executado resultou em estruturas de índices que melhoram a acurácia do sistema CBR tratado de 44.50% para 83.93%. Na medida do nosso conhecimento, é importante ressaltar que este é o primeiro trabalho que busca estudar a relevância de atributos usados em computações de similaridade a partir da utilização de algoritmos de clustering e avaliação de resultados destes algoritmos. Além disso, na medida do conhecimento, ao avaliar o framework proposto e comparar os resultados com resultados disponíveis na literatura, os resultados obtidos na investigação desta dissertação foram melhores que resultados obtidos com outras técnicas nestas mesmas bases de casos. Nesse caso, no domínio de *glass*, o processo executado resultou em estruturas de índices que melhoraram a acurácia do sistema CBR tratado de 75.2% na literatura (DOMENICONI et al., 2000; FERNÁNDEZ-DELGADO et al., 2014; LIU et al., 2012) para 76.35% nesta dissertação. Além disso, no domínio de *hepatitis*, resultados de avaliação do processo executado demonstraram resultados semelhantes aos encontrados na literatura. Embora discussões sobre resultados de acurácia sejam relevantes e tenham sido realizadas nesta dissertação, a definição de estruturas de índices e a investigação da qualidade dos grupos de casos formados com a execução de diferentes algoritmos de clustering também são resultados bastante relevantes, pois podem ser utilizados em diferentes enfoques para o usuário. Além disso, esses resultados propostos complementam a discussão apresentada sobre acurácia destas bases de casos.

Embora artigos ainda possam ser submetidos a partir dos resultados obtidos nesta dissertação, parte da pesquisa desenvolvida foi aceita como um artigo no *31st The European Simulation and Modelling Conference* (*ESM 2017*), denominado *A Case-Based Reasoning And Clustering Framework For The Development Of Intelligent Agents In Simulation Systems* (LUCCA et al., 2017). Este artigo descreveu um framework de integração entre CBR e clustering para apoiar na construção de sistemas de simulação inteligentes. Essa estrutura baseia-se na reutilização de um repositório de experiências de treinamento de simulação armazenadas como casos, em que cada caso é formado por atributos típicos de um problema de simulação. Aproveitando essas experiências passadas para melhorar o realismo dos exercícios de simulação, o framework inova propondo a exploração de técnicas de clustering tanto na construção de funções de similaridade ajustadas quanto na organização de sub-bases de casos, os quais são fundamentais na recuperação eficiente de casos de bases de casos para apoiar a solução de novos problemas de simulação.

Embora a técnica de indexação explorada nesta dissertação tenha permitido melhorar a acurácia do sistema de simulação desenvolvido, visando o desenvolvimento de trabalhos futuros, ainda é possível repetir este tipo de estudo de caso em outros problemas de aplicação (talvez utilizando bases de casos conhecidas e disponíveis na web). Ainda, seria relevante testar o framework proposto em um problema de aplicação onde os casos sejam representados por um grande número de atributos. A ideia seria avaliar se o processo proposto na dissertação funciona nesta situação e também verificar se o processo permite selecionar atributos a serem indexados. Além disso, no estudo de caso desenvolvido, os usuários do domínio de aplicação sendo considerado (principalmente militares interessados em simulações onde os algoritmos construídos estivessem sendo explorados) não puderam participar como gostaríamos do processo de análise e definição de índices comentando os resultados de clustering. Em trabalhos futuros, esse feedback poderia ser explorado de forma a melhorar os resultados do processo de indexação proposto nesta dissertação. Neste sentido, diferentes formas de visualização de resultados dos algoritmos de clustering poderiam ter sido mais exploradas, visando envolver mais usuários na avaliação desses resultados. Mais ainda, trabalhos futuros podem explorar a caracterização de resultados de recuperação de casos de acordo com os grupos gerados em clustering, assim como explorado por mecanismos avançados de busca na web (CARPINETO et al., 2009). Por fim, resultados de acurácia obtidos a partir da aplicação das propostas descritas nesta dissertação podem ser comparados com resultados de acurácia obtidos a partir do emprego de outros enfoques de aprendizado de máquina (por exemplo, enfoques envolvendo técnicas de aprendizado por reforço) visando o ajuste automático de pesos de atributos e consequente construção de melhores funções de similaridade para a resolução de problemas de aplicação.

REFERÊNCIAS

AAMODT, A.; PLAZA, E. Case-Based Reasoning: Foundational Issues, Methodological Variations, and System Approaches. **AI Communications,** v. 7, n. 1, p. 39-59, 1994.

ALSHAMRANI, B. S.; OSMAN, A. H. Investigation of Hepatitis Disease Diagnosis using Different Types of Neural Network Algorithms. **IJCSNS International Journal of Computer Science and Network Security,** v. 17, n. 2, p. 242-246, 2017.

ARMENGOL, E. Classification of melanomas in situ using knowledge discovery with explained case-based reasoning. **Artificial Intelligence in Medicine,** v. 51, p. 93-105, 2011.

ARSHADI, N.; JURISICA, I. Data Mining for Case-Based Reasoning in High-Dimensional Biological Domains. **IEEE Transactions on Knowledge and Data Engineering,** v. 17, n. 8, p. 1127-1137, 2005.

BENNETT, K. P.; BLUE, J. **A Support Vector Machine Approach to Decision Trees**. Neural Networks Proceedings, 1998. IEEE World Congress on Computational Intelligence. The 1998 IEEE International Joint Conference on. Anchorage, AK, USA, USA: IEEE 1998.

BONZANO, A.; CUNNINGHAM, P.; SMYTH, B. Using introspective learning to improve retrieval in CBR: A case study in air traffic control. In: LEAKE, D. B. e PLAZA, E. (Ed.). **Case-Based Reasoning Research and Development: Second International Conference on Case-Based Reasoning, ICCBR-97 Providence, RI, USA, July 25–27, 1997 Proceedings**. Berlin, Heidelberg: Springer Berlin Heidelberg, 1997. p.291-302. ISBN 978-3-540-69238-6.

Breast Cancer Wisconsin. WOLBERG, D. W. H. http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Original): Olvi Mangasarian 1992.

CARPINETO, C. et al. A Survey of Web Clustering Engines. **ACM Computing Surveys,** v. 41, n. 3, p. 38, 2009.

COPERNICUS, U. N. Datasets used for classification: comparison of results. 2017. Disponível em: < http://fizyka.umk.pl/kis-old/projects/datasets.html#Wisconsin >.

DHIVYAPRIYA, P.; SIVAKUMAR, S. Classification of Cancer Dataset in Data Mining Algorithms Using R Tool. **International Journal of Computer Science Trends and Technology (IJCST),** v. 5, n. 1, p. 79-83, 2017.

DOMENICONI, C.; PENG, J.; GUNOPULOS, D. **An adaptive metric for pattern classification**. Advances in Neural Information Processing Systems 13 (NIPS): MIT Press**:** 458-464 p. 2000.

ESTER, M. et al. **A density-based algorithm for discovering clusters a density-based algorithm for discovering clusters in large spatial databases with noise**. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining. Portland, Oregon: AAAI Press**:** 226-231 p. 1996.

EVERITT, B. **Cluster Analysis**. New York: John Wiley & Sons 1974.

FERNÁNDEZ-DELGADO, M. et al. Do we Need Hundreds of Classifiers to Solve Real World Classification Problems? **Journal of Machine Learning Research,** v. 15, n. 1, p. 3133-3181, 2014.

FUKUNAGA, K.; NARENDRA, P. M. A branch and bound algorithm for computing k-nearest neighbors. **IEEE Transactions on Computers,** v. 100, n. 7, p. 750-753, 1975.

FUNG, G. **A Comprehensive Overview of Basic Clustering Algorithms** 2001.

Glass Identification Database. GERMAN, B. http://archive.ics.uci.edu/ml/datasets/Glass+Identification: Vina Spiehler, Ph.D., DABFT, Diagnostic Products Corporation 1987.

HARTIGAN, J. A. **Clustering Algorithms**. New York, NY, USA: John Wiley & Sons, Inc., 1975. ISBN 047135645X.

Hepatitis Domain. G.GONG e CESTNIK, B. http://archive.ics.uci.edu/ml/datasets/Hepatitis 1988.

HONG, T.-P.; LIOU, Y.-L. **Case-Based Reasoning with Feature Clustering**. 7th International Conference on Cognitive Informatics: IEEE 2008.

JAIN, A. K. Data Clustering: 50 Years Beyond K-means. In: DAELEMANS, W.;GOETHALS, B., *et al* (Ed.). **Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2008, Antwerp, Belgium, September 15-19, 2008, Proceedings, Part I**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008. p.3-4. ISBN 978-3-540-87479-9.

JAIN, A. K.; MURTY, M. N.; FLYNN, P. J. Data clustering: a review. **ACM Computing Surveys (CSUR),** v. 31, n. 3, p. 264-323, 1999.

JENG, B. C.; LIANG, T.-P. Fuzzy indexing and retrieval in case-based systems. **Expert Systems with Applications,** v. 8, n. 1, p. 135-142, 1995.

JUNIOR, A. G. L. **Raciocínio Baseado em Casos em Simulação**. 2016. 80 (Graduação). Curso de Graduação em Ciência da Computação, CT - UFSM.

KAINULAINEN, J. **Clustering Algorithms: Basics and Visualization**. Laboratory of Computer and Information Science - Helsinki University of Technology 2002.

KAUFMAN, L.; ROUSSEEUW, P. J. Divisive Analysis (Program DIANA). In: (Ed.). **Finding Groups in Data**: John Wiley & Sons, Inc., 2008. p.253-279. ISBN 9780470316801.

KIM, K.-S.; HAN, I. The cluster-indexing method for case-based reasoning using self-organizing maps and learning vector quantization for bond rating cases. **Expert Systems with Applications,** v. 21, n. 3, p. 147-156, 2001.

KOLODNER, J. L. An introduction to case-based reasoning. **Artificial Intelligence Review,** v. 6, n. 1, p. 3-34, 1992.

LEAKE, D. B., Ed. **Case-Based Reasoning: Experiences, Lessons, and Future Directions**. Menlo Park, CA: AAAI Press/MIT Press, p.436ed. 1996.

LICHMAN, M. UCI Machine Learning Repository. University of California, Irvine, School of Information and Computer Sciences, 2013. Disponível em: < http://archive.ics.uci.edu/ml >. Acesso em: 30/10/2017.

LIU, J. N. K. et al. **Naive Bayesian Classifier Based on Neighborhood Probability**. 14th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems IPMU. Catania, Italy: Springer**:** 112-121 p. 2012.

LUCCA, M. R. B. **Integração de técnicas de raciocínio baseado em casos e agrupamento de dados na construção de sistemas inteligentes**. 2017. (Mestrado). Programa de Pós-Graduação em Ciência da Computação, UFSM, Santa Maria - RS.

LUCCA, M. R. B. et al. **A Case-Based Reasoning And Clustering Framework For The Development Of Intelligent Agents In Simulation Systems**. Manuscript accepted in the 31st The European Simulation and Modelling Conference (ESM 2017). Lisbon, Portugal. 2017

MADHULATHA, T. S. An overview on clustering methods. **IOSR Journal of Engineering,** v. 2, p. 719 - 725, 2012.

MICHALSKI, R. S.; STEPP, R. Automated construction of classifications: Conceptual clustering versus numerical taxonomy. **IEEE Trans. on Pattern Analysis and Machine Intelligence,** v. PAMI-5, n. 4, p. 396-410, 1983.

MITTAL, A.; SHARMA, K. K.; DALAL, S. Applying Clustering algorithm in case retrieval phase of the case-based reasoning. **Apurva Mittal al. International Journal of Research Aspects of Engineering and Management,** v. 1, n. 2, p. 14-16, 2014.

MÜLLER, G.; BERGMANN, R. **A cluster-based approach to improve similarity-based retrieval for process-oriented case-based reasoning**. 20th European Conference on Artificial Intelligence (ECAI 2014). Prague, Czech Republic: IOS Press**:** 639-644 p. 2014.

RICHTER, M. M. The Knowledge Contained in Similarity Measures. Invited Talk at The First International Conference on Case-Based Reasoning (ICCBR'95), 1995. Sesimbra, Portugal.

SILVA, L. A. D. L. **Enhancement of Case-Based Reasoning through Informal Argumentation, Reasoning Templates and Numerical Taxonomy**. 2010. 304 (PhD in Computer Science). Department of Computer Science, University College London, London

SNEATH, P. H.; SOKAL, R. R. **Numerical Taxonomy: The Principles and Practice of Numerical Classification**. San Francisco: W. H. Freeman and Company, 1973.

ŠTER, B.; DOBNIKAR, A. **Neural networks in medical diagnosis: Comparison with other methods**. Proceedings of the International Conference on Engineering Applications of Neural Networks EANN. 91**:** 427-430 p. 1996.

VERNET, D.; GOLOBARDES, E. An Unsupervised Learning Approach for Case-Based Classifier Systems. **Expert Update. The Specialist Group on Artificial Intelligence,** v. 6, n. 2, p. 37-42, 2003.

WEISS, S. M.; KAPOULEAS, I. **An empirical comparison of pattern recognition, neural nets and machine learning classification methods**. Readins in Machine Learning. CA: J.W. Shavlik, TG. Dietterich, Morgan Kauffman 1990.

WETTSCHERECK, D.; AHA, D. W.; MOHRI, T. A Review and Empirical Evaluation of Feature Weighting Methods for a Class of Lazy Learning Algorithms. **Artificial Intelligence Review,** v. 11, n. 1-5, p. 273-314, 1997.

YANG, Q.; WU, J. **Keep It Simple: A Case-Base Maintenance Policy Based on Clustering and Information Theory**. Proceedings of the 13th Biennial Conference of the Canadian Society for Computational Studies of Intelligence. INTELLIGENCE, A. I. A. Montreal, Canada: Springer Berlin Heidelberg. 1822**:** 102-114 p. 2000.

YOUH, H.; RUMBE, G. Comparative Study of Classification Techniques on Breast Cancer FNA Biopsy Data. **International Journal of Interactive Multimedia and Artificial Intelligence,** v. 1, n. A Direct Path to Intelligent Tools, p. 5-12, 2010.

ANEXO A – RESULTADOS DE AVALIAÇÃO DE AGRUPAMENTOS DE CASOS DA BASE DE CASOS DO SIS-ASTROS*

Algoritmo de clustering	Parâmetros de entrada definidos	Métricas	Esquema de avaliação de relevância individual*												
			(b) Valores de pesos para um atributo de caso selecionado = HIGH_VALUE, enquanto todos outros valores de peso para atributos = 1.0												
			(b1) Atr. 1	(b2) Atr. 2	(b3) Atr. 3	(b4) Atr. 4	(b5) Atr. 5	(b6) Atr. 6	(b7) Atr. 7	(b8) Atr. 8	(b9) Atr. 9	(b10) Atr. 10	(b11) Atr. 11	(b12) Atr. 12	(b13) Atr. 13
Densidade – 16 grupos formados	Eps = 3.5 minpts = 15	Pureza	0.29	0.33	0.48	0.39	0.24	0.42	0.38	0.40	0.53	0.49	0.47	0.39	0.36
		Entropia	-354.62	-210.11	-226.96	-245.19	-195.66	-205.56	-263.85	-245.63	-366.87	-322.88	-405.45	-326.80	-340.21
		Precisão	0.46	0.66	0.42	0.42	0.51	0.43	0.44	0.44	0.90	0.75	0.81	0.46	0.51
Hierárquico - Ponto de corte = 15 grupos	Linkage = Complete	Pureza	0.08	0.08	0.14	0.06	0.05	0.07	0.08	0.05	0.39	0.33	0.14	0.08	0.06
		Entropia	-25.89	-28.87	-34.19	-17.17	-15.64	-22.21	-25.50	-16.49	-201.26	-217.03	-43.05	-19.36	-21.28
		Precisão	0.80	0.69	0.58	0.49	0.45	0.61	0.58	0.55	0.94	0.92	0.91	0.61	0.62
	Linkage = Average	Pureza	0.14	0.11	0.11	0.17	0.10	0.08	0.15	0.08	0.29	0.17	0.18	0.15	0.17
		Entropia	-99.14	-82.41	-71.56	-79.56	-65.23	-45.23	-51.54	-57.63	-190.87	-145.60	-93.89	-140.55	-42.22
		Precisão	0.62	0.59	0.61	0.71	0.63	0.60	0.59	0.60	0.90	0.63	0.58	0.71	0.70
	Linkage = Single	Pureza	0.09	0.09	0.11	0.10	0.04	0.06	0.10	0.07	0.22	0.23	0.21	0.11	0.09
		Entropia	-52.10	-27.56	-45.21	-22.75	-19.78	-35.23	-45.11	-17.52	-150.63	-180.34	-165.73	-19.53	-22.85
		Precisão	0.47	0.56	0.55	0.60	0.49	0.62	0.53	0.58	0.85	0.61	0.68	0.71	0.67
Particional	K = 15	Pureza	0.10	0.13	0.14	0.09	0.11	0.16	0.09	0.09	0.25	0.24	0.36	0.09	0.09
		Entropia	-131.22	-167.12	-183.14	-127.92	-130.40	-192.98	-132.89	-136.82	-198.33	-205.09	-221.18	-129.87	-165.99
		Precisão	0.47	0.59	0.64	0.46	0.45	0.53	0.47	0.47	0.71	0.69	0.88	0.47	0.47

* Agrupamentos de Casos formados em clustering onde valores para cada resultado de avaliação de atributo acima do baseline são destacados

* A cor cinza indica valores que são maiores que valores baseline correspondentes.

ANEXO B – RESULTADOS DE AVALIAÇÃO DE AGRUPAMENTOS DE CASOS DA BASE DE CASOS DE *HEPATITIS**

	(b) Valores de pesos para um atributo de caso selecionado = HIGH_VALUE, enquanto todos outros valores de peso para atributos = 1.0																		
	At. 1	At. 2	At. 3	At. 4	At. 5	At. 6	At. 7	At. 8	At. 9	At. 10	At. 11	At. 12	At. 13	At. 14	At. 15	At. 16	At. 17	At. 18	At. 19
Densidade Pureza	0.06	0.07	0.22	0.20	0.18	0.17	0.23	0.29	0.08	0.38	0.12	0.12	0.14	0.31	0.42	0.13	0.33	0.36	0.52
Densidade Entropia	-170.5	-150.8	-220.6	-225.6	-170.5	-214.8	-239.3	-250.6	-177.5	-265.3	-267.5	-165.5	-178.6	-210.6	-280.6	-180.7	-277.2	-225.8	-260.0
Densidade Precisão	0.32	0.24	0.39	0.26	0.37	0.38	0.35	0.24	0.31	0.48	0.46	0.35	0.28	0.33	0.56	0.41	0.54	0.50	0.62
Complete Pureza	0.24	0.20	0.25	0.21	0.23	0.23	0.30	0.32	0.23	0.37	0.52	0.47	0.42	0.60	0.62	0.20	0.28	0.30	0.38
Complete Entropia	-200.8	-205.7	-215.9	-260.3	-217.4	-270.9	-310.2	-350.8	-200.9	-320.4	-330.6	-208.1	-217.0	-231.0	-391.0	-203.6	-270.9	-250.3	-390.6
Complete Precisão	0.35	0.38	0.32	0.35	0.40	0.39	0.44	0.45	0.46	0.49	0.62	0.57	0.50	0.65	0.65	0.32	0.30	0.37	0.42
Single Pureza	0.25	0.18	0.19	0.25	0.31	0.35	0.41	0.44	0.27	0.36	0.42	0.42	0.37	0.46	0.50	0.25	0.24	0.30	0.47
Single Entropia	-221.8	-172.4	-170.6	-221.1	-170.9	-260.7	-278.4	-260.0	-190.9	-298.7	-265.0	-180.9	-210.0	-200.3	-310.4	-240.3	-280.5	-270.9	-270.4
Single Precisão	0.32	0.30	0.29	0.35	0.42	0.44	0.49	0.49	0.36	0.50	0.58	0.59	0.57	0.67	0.50	0.28	0.32	0.38	0.67
Average Pureza	0.17	0.18	0.20	0.17	0.25	0.25	0.32	0.35	0.20	0.31	0.44	0.40	0.38	0.49	0.52	0.29	0.25	0.36	0.55
Average Entropia	-140.3	-152.8	-157.9	-220.6	-180.4	-241.9	-257.4	-290.8	-200.7	-274.2	-273.4	-160.0	-180.1	-202.0	-297.9	-259.1	-240.7	-270.0	-291.5
Average Precisão	0.38	0.34	0.32	0.40	0.31	0.46	0.53	0.54	0.39	0.52	0.60	0.61	0.58	0.63	0.52	0.31	0.34	0.40	0.72
Particional Pureza	0.08	0.07	0.10	0.18	0.12	0.10	0.17	0.23	0.16	0.34	0.32	0.28	0.25	0.38	0.31	0.14	0.10	0.28	0.38
Particional Entropia	-182.6	-171.8	-172.8	-203.1	-174.8	-258.2	-267.8	-284.6	-207.7	-245.0	-253.1	-186.7	-223.4	-218.9	-287.6	-274.5	-247.8	-272.6	-280.5
Particional Precisão	0.22	0.17	0.18	0.23	0.15	0.14	0.19	0.20	0.20	0.24	0.26	0.33	0.38	0.44	0.43	0.10	0.09	0.14	0.33

* Agrupamentos de Casos formados em clustering onde valores para cada resultado de avaliação de atributo acima do baseline são destacados

* A cor cinza indica valores que são maiores que valores baseline correspondentes.

ANEXO C – RESULTADOS DE AVALIAÇÃO DE AGRUPAMENTOS DE CASOS DA BASE DE CASOS DE *GLASS**

Algoritmo de clustering	Parâmetros de entrada definidos	Métricas	(b1) Atr. 1	(b2) Atr. 2	(b3) Atr. 3	(b4) Atr. 4	(b5) Atr. 5	(b6) Atr. 6	(b7) Atr. 7	(b8) Atr. 8	(b9) Atr. 9
Densidade – 7 grupos formados	Eps = 5.5 minpts = 7	Pureza	0.45	0.24	0.28	0.24	0.32	0.41	0.68	0.39	0.70
		Entropia	-140.21	-118.52	-132.32	-110.47	-147.98	-195.51	-220.00	-136.95	-189.36
		Precisão	0.55	0.37	0.48	0.30	0.63	0.74	0.77	0.42	0.45
Hierárquico - Ponto de corte = 7 grupos	Linkage = Complete	Pureza	0.61	0.42	0.41	0.30	0.62	0.70	0.73	0.48	0.57
		Entropia	-132.32	-103.35	-110.23	-105.62	-135.23	-160.78	-190.51	-137.63	-180.80
		Precisão	0.40	0.46	0.50	0.38	0.60	0.71	0.79	0.49	0.73
	Linkage = Average	Pureza	0.55	0.30	0.23	0.24	0.54	0.69	0.67	0.35	0.68
		Entropia	-131.54	-85.65	-90.45	-95.65	-140.52	-162.85	-138.62	-130.47	-165.84
		Precisão	0.64	0.38	0.37	0.36	0.68	0.88	0.59	0.48	0.88
	Linkage = Single	Pureza	0.65	0.30	0.40	0.32	0.37	0.49	0.43	0.31	0.47
		Entropia	-117.56	-90.21	-98.63	-101.68	-131.84	-135.74	-160.64	-127.11	-155.78
		Precisão	0.72	0.47	0.45	0.32	0.59	0.79	0.61	0.50	0.78
Particional	K = 7	Pureza	0.31	0.12	0.10	0.15	0.21	0.34	0.27	0.23	0.34
		Entropia	-170.53	-150.63	-153.44	-140.87	-168.85	-173.52	-180.74	-154.63	-172.51
		Precisão	0.47	0.25	0.32	0.21	0.44	0.43	0.38	0.34	0.39

* Agrupamentos de Casos formados em clustering onde valores para cada resultado de avaliação de atributo acima do baseline são destacados
* A cor cinza indica valores que são maiores que valores baseline correspondentes.

ANEXO D –RESULTADOS DE AVALIAÇÃO DE AGRUPAMENTOS DE CASOS DA BASE DE CASOS DE *BREAST CANCER**

Algoritmo de clustering	Parâmetros de entrada definidos	Métricas	(b1) Atr. 1	(b2) Atr. 2	(b3) Atr. 3	(b4) Atr. 4	(b5) Atr. 5	(b6) Atr. 6	(b7) Atr. 7	(b8) Atr. 8	(b9) Atr. 9
Densidade – 3 grupos formados	Eps = 3.4 minpts = 2	Pureza	0.20	0.22	0.51	0.31	0.23	0.21	0.24	0.47	0.59
		Entropia	-157.54	-221.85	-298.23	-201.72	-185.65	-174.77	-186.74	-241.39	-239.20
		Precisão	0.52	0.48	0.65	0.70	0.44	0.40	0.46	0.69	0.61
Hierárquico - Ponto de corte = 2 grupos	Linkage = Complete	Pureza	0.40	0.40	0.59	0.62	0.23	0.31	0.40	0.53	0.52
		Entropia	-125.65	-147.67	-191.02	-162.10	-121.58	-105.20	-110.54	-138.63	-152.33
		Precisão	0.24	0.29	0.43	0.38	0.20	0.25	0.33	0.31	0.39
	Linkage = Average	Pureza	0.28	0.45	0.44	0.39	0.15	0.25	0.31	0.40	0.40
		Entropia	-137.63	-150.85	-195.80	-178.48	-120.75	-105.21	-115.78	-168.30	-135.41
		Precisão	0.37	0.59	0.55	0.44	0.38	0.36	0.48	0.67	0.64
	Linkage = Single	Pureza	0.39	0.44	0.47	0.47	0.22	0.38	0.32	0.49	0.43
		Entropia	-101.85	-152.22	-139.26	-130.99	-101.50	-84.22	-105.65	-149.55	-109.10
		Precisão	0.46	0.70	0.74	0.73	0.42	0.35	0.42	0.49	0.42
Particional	K = 2	Pureza	0.26	0.33	0.51	0.45	0.34	0.19	0.25	0.32	0.36
		Entropia	-182.54	-190.22	-285.65	-270.98	-160.98	-172.44	-188.74	-239.28	-222.50
		Precisão	0.54	0.62	0.59	0.72	0.50	0.53	0.55	0.74	0.59

* Agrupamentos de Casos formados em clustering onde valores para cada resultado de avaliação de atributo acima do baseline são destacados
* A cor cinza indica valores que são maiores que valores baseline correspondentes.