UNIVERSIDADE FEDERAL DE SANTA MARIA
CENTRO DE TECNOLOGIA
CURSO DE CIÊNCIA DA COMPUTAÇÃO

APLICAÇÃO DE MINERAÇÃO DE DADOS NA AVALIAÇÃO DA RELAÇÃO ENTRE TEMPESTADES GEOMAGNÉTICAS E MUONS

TRABALHO DE GRADUAÇÃO

Miriam Pizzatto Colpo

Santa Maria, RS, Brasil

2011

APLICAÇÃO DE MINERAÇÃO DE DADOS NA AVALIAÇÃO DA RELAÇÃO ENTRE TEMPESTADES GEOMAGNÉTICAS E MUONS

Por

Miriam Pizzatto Colpo

Trabalho de Graduação apresentado ao Curso de Ciência da Computação da Universidade Federal de Santa Maria (USFM, RS), como requisito parcial para a obtenção de grau de
Bacharel em Ciência da Computação

Orientador: Profª. Drª. Lisandra Manzoni Fontoura

Co-orientador: Dr. Adriano Petry

Trabalho de Graduação Nº 336
Santa Maria, RS, Brasil
2011

Universidade Federal de Santa Maria
Centro de Tecnologia
Curso de Ciência da Computação

A Comissão Examinadora, abaixo assinada,
aprova o Trabalho de Graduação

APLICAÇÃO DE MINERAÇÃO DE DADOS NA AVALIAÇÃO DA RELAÇÃO ENTRE TEMPESTADES GEOMAGNÉTICAS E MUONS

elaborado por
Miriam Pizzatto Colpo

como requisito parcial para obtenção do grau de
Bacharel em Ciência da Computação

COMISSÃO EXAMINADORA:

Profª. Drª. Lisandra Manzoni Fontoura
(Presidente/Orientador)

Profª. Drª. Deise de Brum Saccol (UFSM)

Profª. Drª. Marcia Pasin (UFSM)

Santa Maria, 15 de Dezembro de 2011.

“Quem pretende apenas a glória não a merece.”

—MÁRIO QUINTANA

AGRADECIMENTOS

Agradeço primeiramente a Deus, por abençoar-me com a família que tenho, através da qual recebo, incondicionalmente, toda a força e orientação necessárias para seguir minha vida, buscando a realização do que anseio e acredito.

Aos meus pais, Cláudio e Eliane, pela educação, pela confiança e fé que sempre depositaram em mim, pelo constante incentivo e, principalmente, pelo exemplo de caráter que representam. À minha irmã, Danieli, que me acompanhou diariamente durante o período da graduação, pelos conselhos, risos e dificuldades compartilhadas. Ao meu irmão caçula, Miccael, do qual senti saudades diárias durante a graduação (principalmente pelas brigas), por estar sempre na torcida pelo meu melhor.

Aos amigos de todas as fases da minha vida, em especial à Glivia, por fazer-se sempre presente, pelas gargalhadas, conversas e companhia. Aos colegas, também amigos, com quem convivi diariamente durantes esses últimos quatro anos, que tornaram o período da graduação mais fácil e divertido, em especial aos meus queridos Brunão, Dorgas e FF (companheiros de tantos trabalhos, desesperos e risadas); Grahl e Reis (colegas de turma e INPE, companhias garantidas para cafés, ajuda em debugs e risos); Evandro, Nêne, Proibidão, Porto, Fabri, Nathan e "lacaboys" (Bernardo, Fred, Guilherme e Cícero).

Meus sinceros agradecimentos também a todos os professores que de alguma forma já contribuíram para o meu crescimento. Ao professor Fábio, da Escola Estadual de Educação Básica Manoel Viana, por ter aumentado, através das suas aulas, o meu gosto pelas exatas. Ao professor Marcelo, da UFSM, por acompanhar nossa turma nos semestres iniciais do curso, sempre se mostrando preocupado com o nosso entendimento e apresentando as aplicações dos Cálculos na nossa área, fazendo com que tudo parecesse fácil, com suas várias explicações e notável gosto pelo que faz. À professora Lisandra pela ajuda durante as várias disciplinas ministradas durante o curso, por mostrar-se sempre acessível e disposta a compartilhar seus conhecimentos e experiências, e por aceitar orientar este trabalho. À professora Andrea pelos ensinamentos e pela indicação à iniciação científica no INPE, da qual resultou este trabalho. Ao Dr. Adriano, por dar-me a oportunidade de participar de sua equipe e compartilhar seus conhecimentos, pela atenção, ajuda e orientação neste trabalho, como em todas as atividades nas quais estive envolvida no INPE. E, por fim, às professoras Deise e Marcia pelos ensinamentos e pela disponibilidade de fazerem parte da banca examinadora desse trabalho.

RESUMO

Trabalho de Graduação

Curso de Ciência da Computação

Universidade Federal de Santa Maria

APLICAÇÃO DE MINERAÇÃO DE DADOS NA AVALIAÇÃO DA RELAÇÃO ENTRE TEMPESTADES GEOMAGNÉTICAS E MUONS

Autor: Miriam Pizzatto Colpo
Orientador: Profª. Drª. Lisandra Manzoni Fontoura
Co-orientador: Dr. Adriano Petry
Local e data da defesa: Santa Maria, 15 de dezembro de 2011.

As tempestades geomagnéticas vêm sendo um objeto de estudo recorrente na área de Clima Espacial, devido às consequências que elas, em grandes intensidades, podem trazer para a superfície terrestre. Ejeções de massa coronal são potenciais causadoras desses fenômenos e podem ocasionar relevantes variações na intensidade de raios cósmicos secundários, tais como os muons. O Instituto Nacional de Pesquisas Espaciais participa da Rede Global de Detectores de Muons através do Detector Multidirecional de Muons (MMD), instalado junto ao Observatório Espacial do Sul, capaz de efetuar contagens da incidência de muons na atmosfera terrestre em vários canais direcionais. Este trabalho tem como objetivo a aplicação de algoritmos de mineração de dados aos dados de contagem do MMD a fim de possibilitar a descoberta de padrões, verificando a associatividade dos muons às tempestades geomagnéticas.

Palavras-chave: Muons, Tempestades Geomagnéticas, Mineração de Dados.

ABSTRACT

Undergraduate Final Work

Graduation in Computer Science

Federal University of Santa Maria

APPLICATION OF DATA MINING IN EVALUATING OF THE RELATION BETWEEN GEOMAGNETIC STORMS AND MUONS

Author: Miriam Pizzatto Colpo
Adviser: Profª. Drª. Lisandra Manzoni Fontoura
Co-adviser: Dr. Adriano Petry

The geomagnetic storms have been a recurring subject of study in the area of Space Weather, due to the consequences that they, in large amounts, can bring to the Earth's surface. Coronal Mass Ejections are potential causes of these phenomena and can cause significant variations in the intensity of secondary cosmic rays, such as muons. The National Institute for Space Research participates of the Global Muon Detector Network by a Multidirectional Muon Detector (MMD), located at the Southern Space Observatory. This instrument performs the counting of muons incidence in the atmosphere at various directional channels. This study aims to apply data mining algorithms to count data from the MMD to enable the discovery of patterns, verifying the associativity of the muons to geomagnetic.

Keywords: Muons, Geomagnetic Storms, Data Mining.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ARFF	Attribute-Relation File Format
CMEs	Coronal Mass Ejections
CRISP-DM	Cross-Industry Standard Process for Data Mining
Dst	Disturbance Storm-Time
DTM	Decision Table Majority
FPGA	Field Programmable Gate Array
GMDN	Global Muon Detector Network
GNU	General Public License
GPS	Sistema de Posicionamento Global
ICMEs	Interplanetary Coronal Mass Ejections
IMF	Interplanetary Magnetic Field
INPE	Instituto Nacional de Pesquisas Espaciais
iSWA	integrated Space Weather Analysis System
KDD	Knowledge Discovery in Database
MMD	Multidirectional Muon Detector
MMD-DB	Multidirectional Muons Detector Database
OES	Observatório Espacial do Sul
UFSM	Universidade Federal de Santa Maria
WDC	World Data Center for Geomagnetism
WEKA	Waikato Environment for Knowledge Analysis

SUMÁRIO

1 INTRODUÇÃO

Segundo EMBRACE (2011), o Clima Espacial é a área de conhecimento dos fenômenos solares e suas ocorrências, que se manifestam de forma recorrente e afetam os astros e artefatos no espaço. Dentre os fatores importantes ao Clima Espacial, encontram-se o conhecimento e previsão de fenômenos que afetam de forma direta as atividades humanas, tais como atividades solares e tempestades geomagnéticas.

As tempestades geomagnéticas são distúrbios no campo magnético da Terra, que ao ocorrerem, dependendo da intensidade, podem causar danos no espaço e na superfície terrestre, principalmente em sistemas tecnológicos, tais como telecomunicação, energia elétrica e satélites. A fim de amenizar os possíveis efeitos ocasionados por esse fenômeno, existe um esforço em pesquisas da área de Clima Espacial para prevê-lo, o que permitiria que ações fossem tomadas para garantir a integridade dos sistemas, dado um alerta de ocorrência. Dentre as potenciais causas das tempestades geomagnéticas estão as ejeções de massa coronal interplanetárias (Moldwin, 2008), que são estruturas oriundas do sol que podem causar relevantes variações na intensidade de raios cósmicos primários, partículas de alta energia que ao colidirem com a atmosfera terrestre dão origem aos muons (raios cósmicos secundários).

A relação entre as tempestades geomagnéticas e a variação da intensidade de raios cósmicos primários, e, consequentemente, de muons, vêm motivando o estudo dos dados de incidência desses raios cósmicos secundários, o que explica a existência de uma rede global de detectores de muons, que conta com a colaboração de dez instituições de seis países. A Universidade Federal de Santa Maria (UFSM) e o Instituto Nacional de Pesquisas Espaciais (INPE) fazem parte dessa rede, representando o Brasil, através de um detector instalado junto ao Observatório Espacial do Sul (OES/CRS/INPE-MCT), em São Martinho da Serra – RS, responsável por coletar dados de contagens da incidência de muons em vários canais direcionais com intervalos de freqüência de um minuto e dez minutos, gerando um volume de dados considerável.

A existência de grandes volumes de dados, com possíveis padrões implícitos, permite a extração de novos conhecimentos, uma tarefa humanamente impossível que pode ser realizada por meio da mineração de dados. A mineração de dados é uma forma de detectar padrões, relações, regras e associações a partir de dados brutos, permitindo a extração de conhecimentos úteis a partir de grandes volumes de informações (LAROSE, 2005). A mineração faz uso de técnicas e algoritmos com características multidisciplinares, fundamentados em áreas como análise estatística, banco de dados e inteligência artificial (BERNARDI, 2010) e, embora tenha surgido no âmbito comercial, vem sendo aplicada também no domínio científico.

1.1 Objetivos

1.1.1 Objetivo Geral

O objetivo geral deste trabalho consiste em aplicar algoritmos de mineração de dados, com o apoio de uma ferramenta de mineração, aos dados de incidência de muons, oriundos do detector brasileiro, a fim de verificar a associatividade desses raios cósmicos secundários às tempestades geomagnéticas.

1.1.2 Objetivos Específicos

- Conhecer a estrutura dos dados disponibilizados pelo detector de muons brasileiro;
- Definição e conhecimento de uma ferramenta de mineração de dados para ser usada;
- Analisar e escolher a técnica e os algoritmos de mineração mais adequados ao caso específico dos muons;
- Preparar os dados, incluindo:
    - Buscar e escolher dados relevantes, que indiquem a ocorrência de tempestades geomagnéticas;
    - Desenvolver um *software* que colete as informações escolhidas das fontes descobertas e as armazene em um arquivo com formato específico de entrada da ferramenta de mineração escolhida;

- Aplicar os algoritmos de mineração aos dados previamente preparados e analisar os resultados.

1.2 Estrutura do Texto

Este trabalho está organizado da seguinte maneira: no Capítulo 2 é descrita uma fundamentação teórica sobre os assuntos abordados no trabalho, incluindo os conceitos físicos acerca dos muons e das tempestades geomagnéticas e uma revisão bibliográfica de mineração de dados. No Capítulo 3 é explicado como o trabalho foi desenvolvido, apresentando a ferramenta escolhida, a preparação dos dados, a técnica e os algoritmos considerados adequados para o domínio em questão. No Capítulo 4 são apresentados os testes e os resultados obtidos através da aplicação dos algoritmos de mineração de dados. E, finalmente, no Capítulo 5 são descritas as conclusões deste trabalho.

2 FUNDAMENTAÇÃO TEÓRICA

Este capítulo trata dos conceitos físicos envolvendo a relação Sol-Terra (Clima Espacial), incluindo as tempestades geomagnéticas e os muons, além dos conceitos teóricos acerca da mineração de dados, a fim de proporcionar a devida compreensão dos fundamentos necessários no desenvolvimento desse trabalho.

2.1 Clima Espacial

Segundo MURALIKRISHMA (2009), o Clima Espacial é a área que estuda os processos físicos envolvidos na influência que o Sol exerce nos planetas do meio interplanetário, as causas e as consequências dessa interação, sendo sua variabilidade causada principalmente pela influência exercida sobre os planetas mais próximos do Sol, como a Terra. Na Figura 2.1 são apresentados os três principais fenômenos partidos do Sol que podem resumir a relação Sol-Terra (DAL LAGO, 2003), interessando para este trabalho apenas o último fenômeno, causador das tempestades geomagnéticas.

2.1.1 O Sol

O Sol é uma estrela da Via Láctea de cerca de 4.5 bilhões de anos, constituída principalmente pelos elementos hidrogênio (92.1%), hélio (7.8%), oxigênio (0.061%), carbono (0.030%) e nitrogênio (0.0084%). Ele contém mais de 99% da massa total do sistema solar e tem sua atmosfera, por convenção, dividida em três principais camadas, representadas na Figura 2.2, que são: fotosfera, que é a camada visível a olho nu, cromosfera, acima da primeira, e coroa, que é a camada mais externa da atmosfera solar (MOLDWIN, 2008).

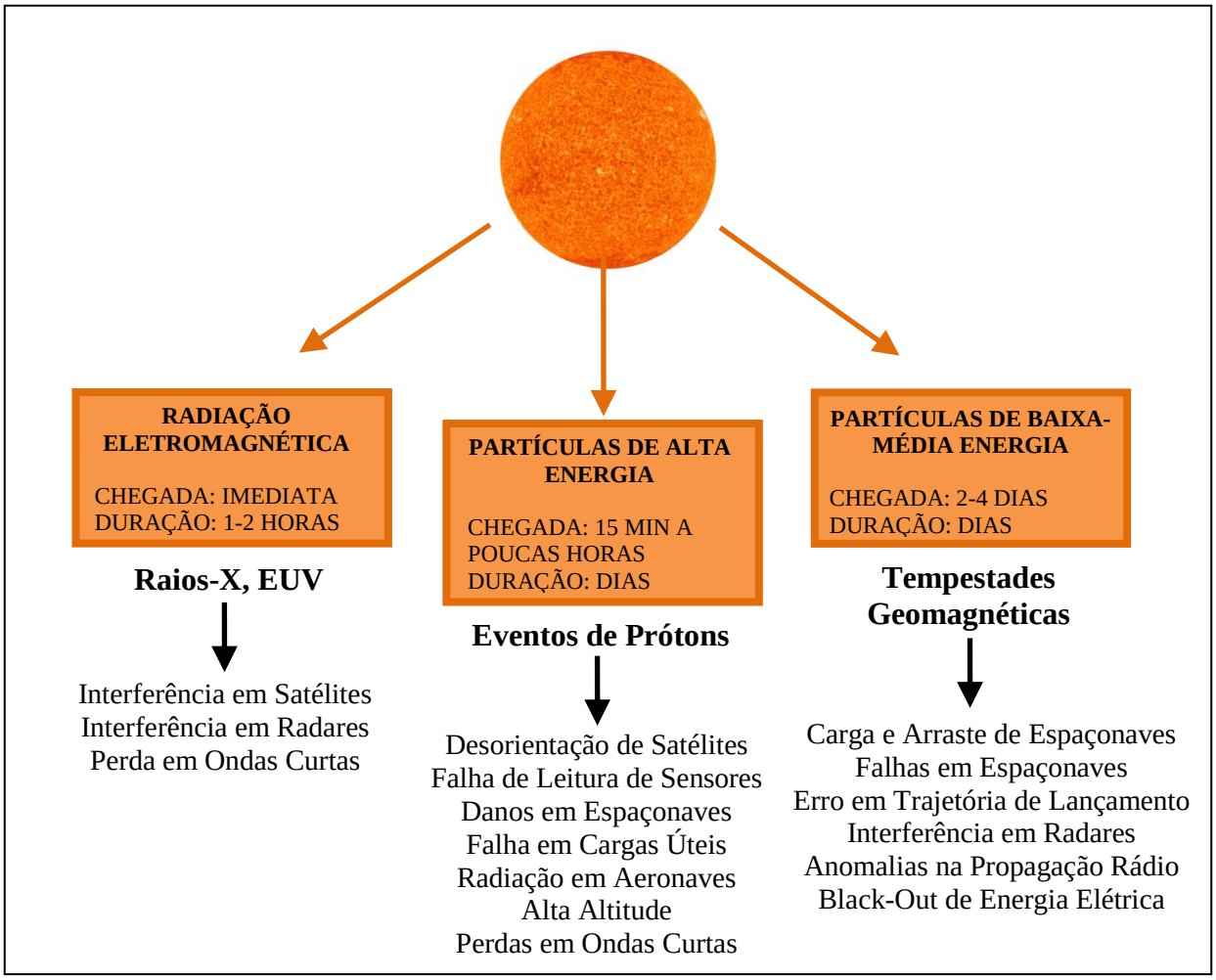


Figura 2.1. Diagrama dos principais fenômenos que constituem o Clima Espacial (Adaptado de: DAL LAGO, 2003).

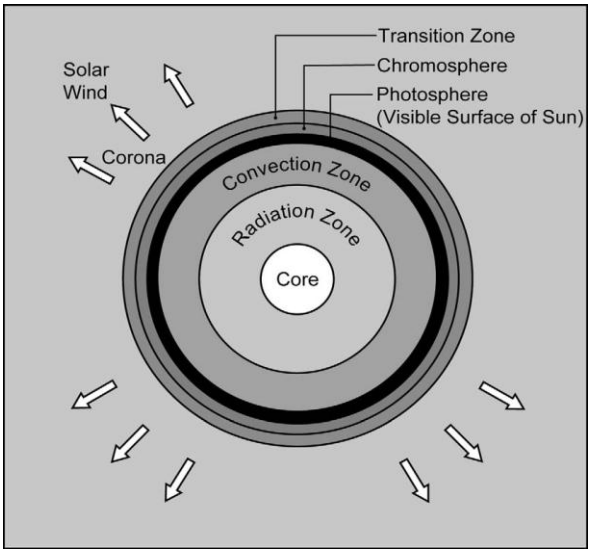


Figura 2.2. As principais regiões do sol. As regiões dentro do Sol são definidas pela forma como a energia é transferida do núcleo para a superfície. As regiões da atmosfera do Sol são definidas por sua densidade e temperatura (MOLDWIN, 2008).

2.1.1.1 A Atividade Solar

Segundo MURALIKRISHNA (2009) “o Sol é um corpo gasoso que está em constante atividade, a qual envolve processos físicos que se iniciam no núcleo e resultam em eventos que podem ser observados nas camadas externas, como na fotosfera”. A variabilidade solar é caracterizada pela mudança no número de manchas solares (*sunspots number*), que são regiões mais frias e escuras que as áreas circunvizinhas, observadas na parte visível do Sol, como mostrado na Figura 2.3.

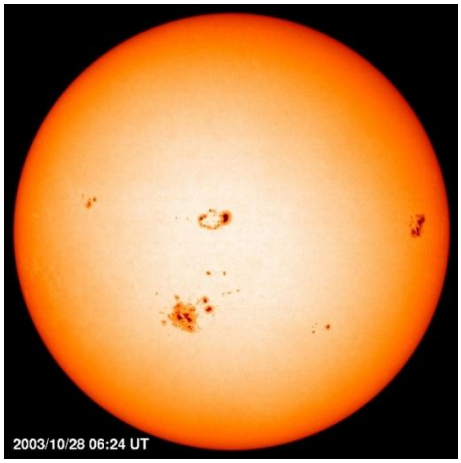


Figura 2.3. Manchas solares registradas em 28 de outubro de 2003 (SOHO, 2011).

As observações dos registros de manchas solares mostram um ciclo regular médio de aproximadamente 11 anos da atividade solar (EDDY, 1976 apud ECHER et al.,2003), ou seja, a cada 11 anos o Sol passa por um período de mínima e outro de máxima atividade magnética. Durante o máximo solar, ocorre um grande aumento na ocorrência de fenômenos energéticos nas regiões associadas às manchas solares (regiões ativas). Estes fenômenos podem ser chamados de explosões ou *flares* solares, caracterizados pela emissão, em curtos intervalos de tempo (variando de alguns segundos até poucas horas, para os fenômenos mais intensos), de grandes quantidades de energia e podem estar relacionadas às ejeções de massa coronal e tempestades magnéticas na Terra (MILANE et al., 2003), que são explicadas a seguir nas Subseções 2.1.1.2 e 2.1.3, respectivamente.

2.1.1.2 Vento solar e CMEs

Segundo MILONE et al. (2003), o vento solar é um fluxo de elétrons e íons positivos expulsos do sol em alta velocidade, que se propaga pelo meio interplanetário. O campo magnético do Sol não se limita apenas à sua vizinhança e uma parte dele é transportada em

direção ao meio interplanetário através do vento solar, o que recebe o nome de Campo Magnético Interplanetário (Interplanetary Magnetic Field - IMF) (MOLDWIN, 2008).

Junto consigo, o vento solar pode transportar também estruturas solares, como as ejeções de massa coronal (do inglês: Coronal Mass Ejections - CMEs), que podem ser associadas à ocorrência de tempestades magnéticas na Terra (MURALIKRISHNA, 2009). As CMEs são "grandes quantidades de matéria, entremeadas de linhas de campo magnético, que são expulsas do Sol durante um período de várias horas, formando uma enorme erupção que se expande para o espaço exterior a velocidades de várias centenas a poucos milhares de km/s" (MILONE et al., 2003) e sua freqüência varia de acordo com o ciclo de atividade solar, aumentando o número de eventos no período de máxima atividade solar. As CMEs podem ter suas características alteradas ao atravessarem o meio interplanetário e, para essa diferenciação, recebem o nome de ICMEs (do inglês: Interplanetary Coronal Mass Ejection) ao serem lançadas no meio interplanetário.

2.1.2 Campo Geomagnético

O campo geomagnético (ou campo magnético terrestre) pode ser, próximo à superfície, aproximado a um dipolo não coincidente com o eixo de rotação, como ilustrado na Figura 2.4, e tem a forma de uma barra magnética, tendo o Pólo Norte Magnético (na região ártica do Canadá) e o Pólo Sul Magnético (no sul da Austrália) (DAL POZ; CAMARGO, 2006).

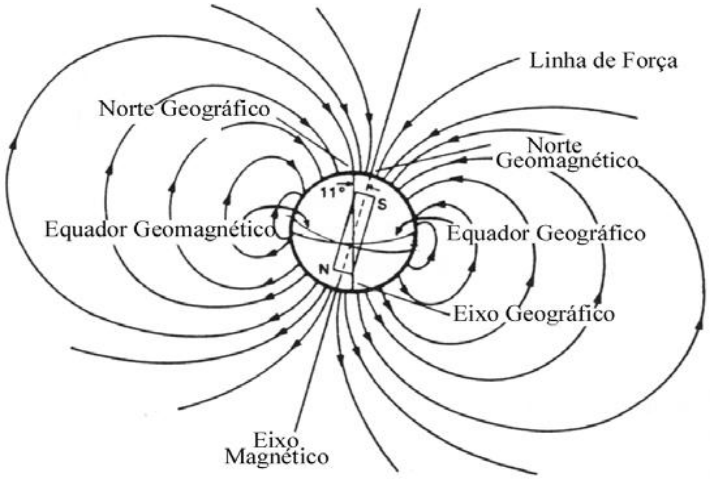


Figura 2.4. Modelo do campo geomagnético (DAL POZ; CAMARGO, 2006).

O campo geomagnético exerce grande influência na variação da densidade de elétrons, tendo suas perturbações refletidas em modificações nas condições de transporte do meio

ionizado, e tem suas linhas de força comprimidas por meio da ação do vento solar (MATSUOKA et al., 2010). O vento solar faz com que o campo magnético se confine e distorça formando a magnetosfera, uma cavidade com uma cauda longa que se estende por vários raios terrestres na direção anti-solar (FEDRIZZI, 2003 apud MATSUOKA et al., 2010). Durante eventos solares (como explosões e CMEs), o vento solar tem seus parâmetros (velocidade e densidade) alterados, gerando uma alteração no campo geomagnético, o que contribui para a ocorrência de tempestades geomagnéticas (MATSUOKA et al., 2010).

2.1.3 Tempestades Geomagnéticas

Observações baseadas em magnetômetros terrestres, feitas na metade do século XIX, constataram fortes flutuações no campo magnético terrestre, sendo denominadas de tempestades geomagnéticas (GONZALEZ et al., 1994 apud SAVIAN et at., 2005). Nas tempestades geomagnéticas uma grande quantidade de energia é transferida do vento solar para dentro da magnetosfera terrestre, o que intensifica as correntes elétricas na magnetosfera e na superfície terrestre. Dentre os efeitos mais conhecidos dessas tempestades estão os diversos prejuízos em satélites, causando danos no Sistema de Posicionamento Global (GPS), em telecomunicações e, até mesmo, em astronautas que se encontram em naves espaciais devido a alta radiação emitida (SAVIAN et al., 2005).

2.1.3.1 Índice Dst

A intensidade de tempestades geomagnéticas pode ser especificada por meio de índices geomagnéticos, sendo as características e a latitude de ocorrência desses distúrbios fatores influentes na escolha do índice mais apropriado (TASCIONE, 1994 apud MURALIKRISHNA, 2009).

O índice Dst (*Disturbance Storm-Time*) "representa o invólucro das curvas de medidas magnetométricas obtidas por uma cadeia de magnetômetros localizadas na região equatorial do globo terrestre" (YAMASHITA, 1999) e é considerado como o índice que melhor define a tempestade geomagnética, por apresentar um comportamento padrão antes e após ocorrência de uma tempestade. Antes do início de uma tempestade, o índice apresenta um pico de intensidade, conhecido como "fase inicial", que é seguido pelo desenvolvimento da "fase principal", caracterizada pela queda brusca do valor do índice. Após alcançar seu mínimo, o índice inicia a recuperação de seu valor até atingir um valor calmo (que não caracteriza uma tempestade), fase conhecida como "fase de recuperação" (YAMASHITA, 1999). Essas fases

são representadas na Figura 2.5, através dos dados de Dst entre os dias 5 e 8 de setembro de 1982, quando ocorreu uma tempestade intensa, vide Tabela 2.1.

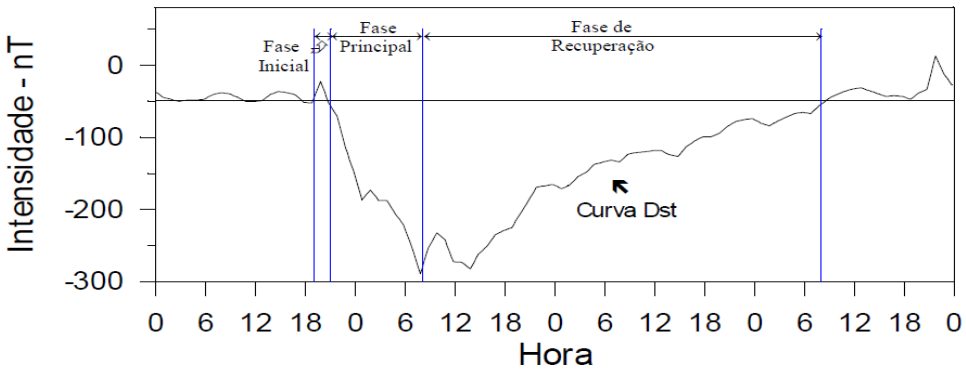


Figura 2.5. Curva Dst entre os dias 5-8 setembro de 1982 com uma intensa tempestade magnética (YAMASHITA, 1999).

Tabela 2.1. Classificação das tempestades geomagnéticas pelo Dst (Adaptada de FEDRIZZI, 2003 apud MATSUOKA, 2010).

Intensidade da Tempestade	**Dst (nT)**
Inexistente	Maiores que -30
Fraca	De -30 a -50
Moderada	De -50 a -100
Intensa	De -100 a -250
Muito intensa	Menores que -250

2.1.4 Raios Cósmicos

A Terra é constantemente bombardeada por átomos altamente ionizados e outras partículas subatômicas, conhecidas como raios cósmicos, que viajam com uma velocidade próxima a da luz e, na maioria, são núcleos de átomos. Embora sejam chamados de raios, os raios cósmicos consistem em partículas energéticas, que podem se originar fora da heliosfera (raios cósmicos galácticos) ou se originar do Sol (partículas energéticas solares). Os raios cósmicos podem ser compostos por qualquer elemento, além de incluírem elétrons, pósitrons (essencialmente um elétron com carga positiva) e outras partículas subatômicas. Por serem partículas carregadas, os raios cósmicos tem seu movimento desviado pelos campos

magnéticos galácticos ao se propagarem pelo espaço interestelar, sendo espalhados em todas as direções (MOLDWIN, 2008).

Ao atingirem a atmosfera terrestre os raios cósmicos de alta energia colidem com partículas atmosféricas, gerando chuvas de partículas secundárias que atingem a superfície. A criação de píons, partículas subatômicas incomuns, é um subproduto dessas colisões, que decaem rapidamente e produzem muons, neutrinos e raios gama (MOLDWIN, 2008).

2.2 Os Muons e a GMDN

Os muons, como mencionado na Seção 2.1.4, são partículas secundárias resultantes da colisão inelástica de partículas primárias de alta energia com partículas da atmosfera terrestre, que mantém a direção e o sentido do raio cósmico primário que os originou. Os muons são partículas com alto poder de penetração e massa aproximadamente 210 vezes maior que a do elétron, com tempo de vida (em repouso) aproximado a $2x10^{-6}$ s, atingindo a superfície terrestre devido suas velocidades relativísticas (SILVA, 2005).

As estruturas solares que se propagam no meio interplanetário, tais como as CMEs, afetam a população de raios cósmicos galácticos pré-existentes de diversas formas, como no decréscimo de Forbush, que é a diminuição da contagem de raios cósmicos observados na superfície durante distúrbios geomagnéticos (SAVIAN, 2005), podendo esse decréscimo ser refletido na incidência de muons.

A Rede Global de Detectores de Muons (do inglês: Global Muon Detector Network – GMDN) é uma colaboração de dez instituições de seis países, que tem como objetivo usar os dados de contagens direcionais da incidência de muons obtidos por quatro detectores de países diferentes para estudos da previsão de estruturas solares potenciais causadoras de tempestades geomagnéticas.

2.2.1 Multidirectional Muon Detector (MMD)

A incidência e a direção dos muons podem ser obtidas por meio da medição de cintilações em um conjunto de detectores dispostos em duas camadas, separadas por uma placa de chumbo de 5cm de largura. Por serem partículas de alta energia, os muons, geralmente, são capazes de atravessar essa placa, sendo os sentidos das incidências obtidos

pela análise da correlação entre a detecção de cintilações nas partes superior e inferior, como ilustrado na Figura 2.6 (PETRY et al., 2011).

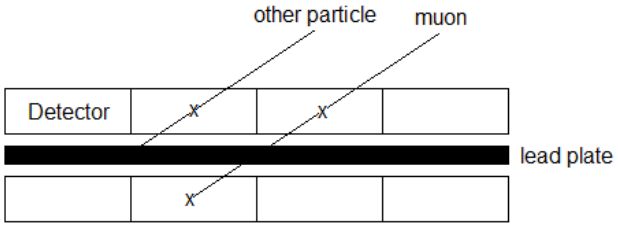


Figura 2.6. Esquema da detecção direcional de muons (PETRY et al., 2011).

O Instituto Nacional de Pesquisas Espaciais (INPE) e a Universidade Federal de Santa Maria (UFSM) são as duas instituições brasileiras a colaborarem com a GMDN, através do Detector Multidirecional de Muons (do inglês: Multidirectional Muon Detector - MMD), instalado no Observatório Espacial do Sul, OES/CRS/INPE – MCT, localizado em São Martinho da Serra no Rio Grande do Sul.

O MMD brasileiro teve sua instalação realizada em 2001, em uma parceria INPE-UFSM, através da cooperação Brasil – Japão – EUA em Clima Espacial. O aparelho era constituído inicialmente por quatro detectores dispostos na camada superior e outros quatro na camada inferior, fornecendo informações em nove canais direcionais (V, N, S, E, W, NE, NW e SE, SW). Após uma atualização, realizada em 2005, o MMD passou a contar com duas camadas de 28 detectores, dando início a operação do sistema de medição em treze canais direcionais (com o acréscimo de N2, S2, E2, W2). Porém, o aumento de detectores poderia fornecer informações em várias outras direções, o que foi possível em 2006, com a instalação de um sistema de captura por FPGA (*Field Programmable Gate Array*) (PETRY, 2010).

2.2.1.1 Sistema de captura por FPGA

O aumento no número de detectores tornou possível mensurar 91 canais direcionais ao invés dos 13 tradicionais. Por isso, desde 2006, outro sistema de medição, baseado em hardwares específicos (FPGAs) está operando simultaneamente, considerando todas as possíveis direções. O sistema opera fazendo medições para 119 canais, pois pretende-se aumentar o número de detectores para 36 por camada (em uma grade de quatro linhas por nove colunas), sendo que as direções ainda não contempladas pelos 28 detectores por camada atuais recebem valor igual a zero (PETRY, 2010).

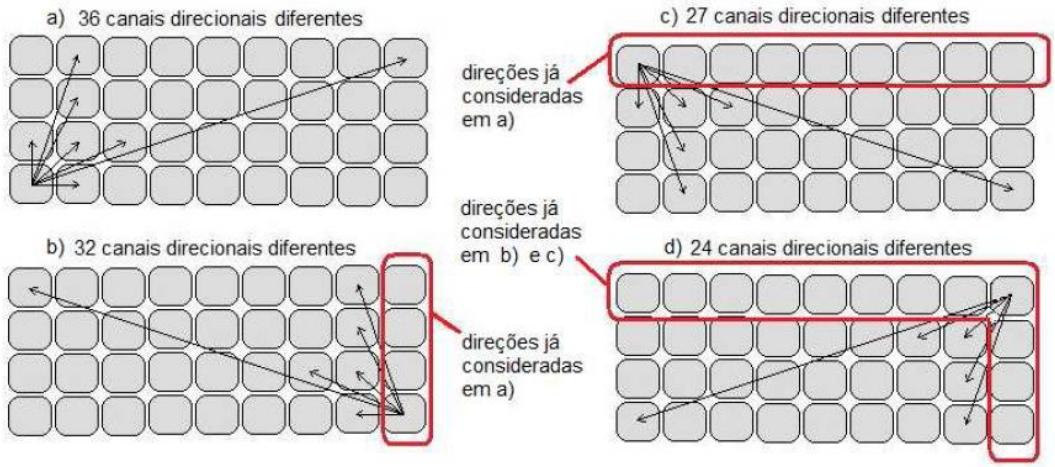


Figura 2.7. Canais direcionais possíveis no MMD composto por 36 detectores em cada camada (PETRY, 2010).

A Figura 2.7 ilustra os possíveis canais direcionais para o novo sistema de contagem, mostrando uma visão superior do MMD, onde as flechas indicam as possíveis direções combinando um detector da camada superior com um da inferior, que não aparece na imagem. Os canais são representados por coordenadas (x, y), que indicam a proximidade entre os detectores da camada superior e da camada inferior, sendo que o valor de "y" pode variar de -3 a 3 e o valor de "x" de -8 a 8, para o instrumento com 36 detectores por camada. Assim, a direção vertical (V) é identificada pela coordenada (0, 0), levando em consideração 36 pares de detectores e a direção norte (N) pela coordenada (0, 1), considerando as contagens em 27 pares de detectores, por exemplo. Na configuração atual (com 28 detectores por camada), para coordenadas com "x" igual a -8, -7, 7 e 8, é atribuído o valor zero (PETRY, 2010).

2.2.1.2 Informações produzidas

Os dados produzidos pelo MMD são gerados e armazenados em arquivos texto em um computador instalado no OES, que auxilia o funcionamento do instrumento. Eles são coletados com intervalos de tempo de um e dez minutos, sendo o registro feito em arquivos diferentes, um para cada resolução temporal. Os arquivos de um e dez minutos contêm, respectivamente, 60 e 144 linhas cada, ou seja, uma hora de operação do instrumento para o primeiro e um dia para o segundo. Esses arquivos são arranjados em subpastas, uma para cada mês, podendo os dados ser acessados remotamente. Os arquivos de dados gerados pelo MMD são, periodicamente, enviados para pesquisadores da Universidade de Shinshu, no Japão, que os disponibiliza on-line em ftp://ftp.bartol.udel.edu/takao/muon_data/ (PETRY et al., 2011). Os dados publicados on-line apresentam-se também por meio de informações horárias

(calculadas a partir dos dados de dez minutos), onde existe um arquivo diário, composto por 24 linhas (cada uma representando uma hora de contagens) (PETRY, 2010).

2.2.1.3 Banco de Dados

Todas as informações coletadas pelo MMD a partir de dezembro de 2006 são armazenadas em um banco de dados relacional, cuja implementação (incluindo modelagem e construção do banco de dados, além do desenvolvimento de um software de carregamento dos dados para o banco) foi realizada pelo Dr. Adriano Petry, tecnologista do INPE, no ano de 2010.

O MMD-DB (do inglês: *Multidirectional Muon Detector Database*) é composto por oito tabelas relacionadas de forma um-para-muitos (1:N), como mostrado na Figura 2.8. As tabelas "instrument_type", "instrument" e "political_location" servem para identificar o instrumento, possibilitando a inclusão de outros detectores além do MMD do OES e as demais tabelas são exclusivamente usadas para o armazenamento dos dados coletados pelo instrumento.

Cada arquivo com dados de medições equivale a um registro na tabela "muon_file", identificando a origem dos dados e cada linha desses arquivos representará um registro da tabela "muon_data" e diversos registros associados (um para cada direção) na tabela "muon_directional_data". Os dados de contagem da incidência de muons podem ser influenciados pela pressão atmosférica no momento da medição, existindo coeficientes barométricos (calculados individualmente para cada canal direcional) para calibrá-los. Esses coeficientes são armazenados na tabela "normalization_coefficients", podendo cada registro da tabela "directions" ser associado a vários registros da "normalization_coefficients", visto que a normalização das contagens é feita individualmente para cada canal direcional, e várias normalizações podem ser identificadas, emperíodos de tempo distintos. O campo "pressure_corrected" da tabela "muon_directional_data" identifica se a contagem associada está ou não normalizada, visto que alguns dados oriundos de arquivos com informações horárias apresentam contagens direcionais já normalizadas barometricamente (PETRY, 2010).

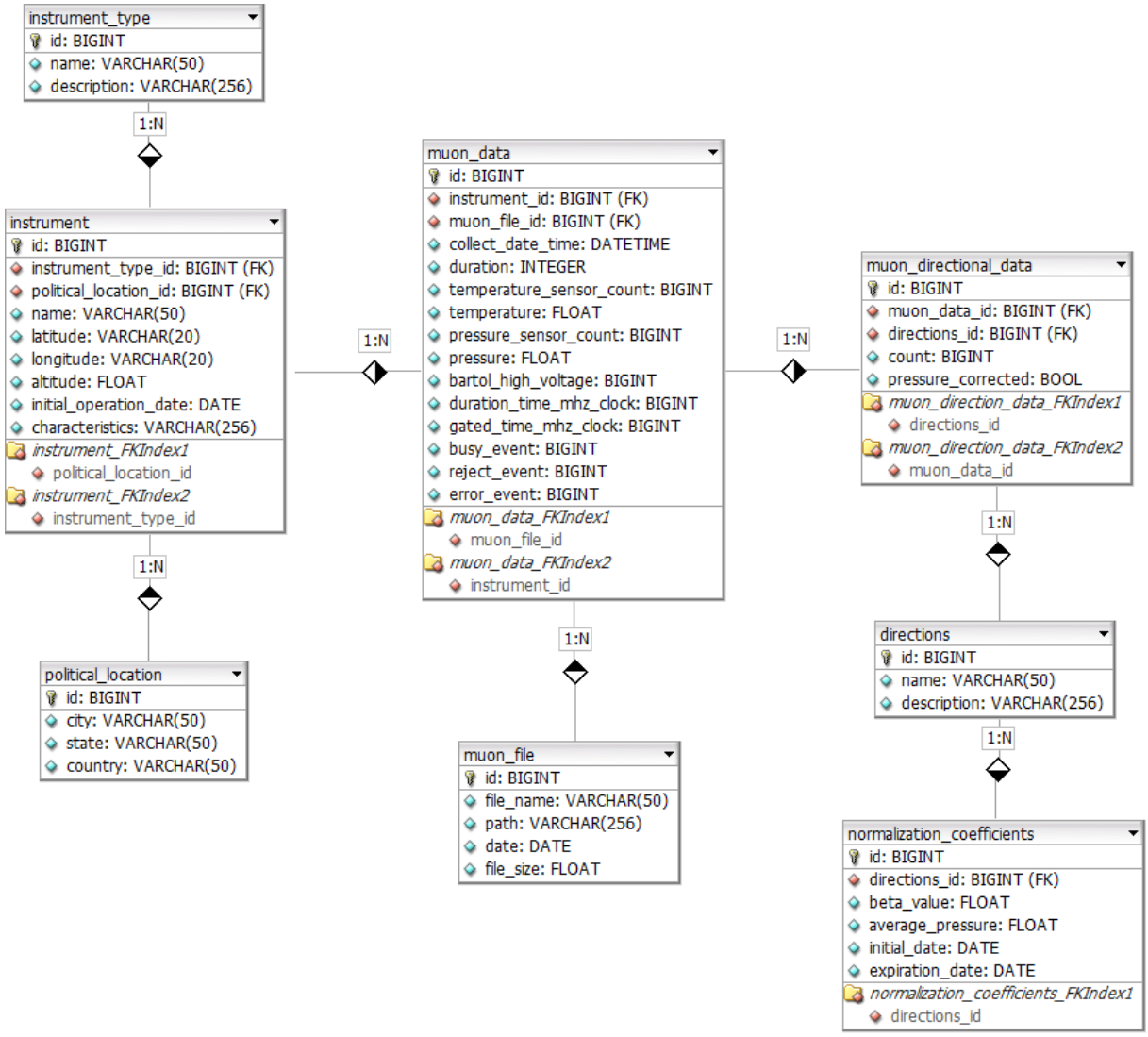


Figura 2.8. Modelagem para o banco de dados do MMD (PETRY et al., 2011).

2.3 Mineração de Dados

Estima-se que a quantidade de informações armazenadas em banco de dados no mundo dobra a cada 20 meses (WITTEN et al., 2011). O rápido crescimento desses volumes de dados excedeu em muito a capacidade humana de compreensão, sendo que a abundância de dados sem a disponibilidade de ferramentas poderosas de análise caracteriza uma situação de “dados ricos, mas informações pobres” (HAN; KAMBER, 2006). Essa situação faz com que os dados de grandes repositórios, que poderiam ajudar na tomada de importantes decisões dentro das corporações, sejam raramente acessados, o que pode ser revertido através do uso de ferramentas de mineração de dados, que objetivam extrair conhecimentos úteis embutidos a esses grandes repositórios (HAN; KAMBER, 2006).

Dentre as definições de mineração de dados, encontram-se:

“Mineração de dados é a análise de (muitas vezes grandes) conjuntos de dados observacionais a fim de encontrar relações insuspeitas e resumir os dados em novas formas que são compreensíveis e úteis para o proprietário dos dados” (HAND et al., 2001).

*“*A mineração provê um método automático para descobrir padrões em dados, sem a tendenciosidade e a limitação de uma análise baseada meramente na intuição humana*”* (BRAGA, 2005).

Para WITTEN et al. (2011), a mineração de dados pode ser definida como um processo automático ou (mais geralmente) semiautomático de descoberta de padrões em dados, devendo os dados existirem em quantidades substanciais e os padrões encontrados possibilitar algumas vantagens (geralmente econômicas).

Segundo KORTH et al. (2006), a mineração distingue-se da descoberta de conhecimento na inteligência artificial (também chamada de aprendizado de máquina) ou na análise estatística apenas por lidar com grandes volumes de dados, armazenados principalmente em bancos de dados.

A mineração de dados foi considerada um desenvolvimento com tendência revolucionária, pela ZDNET *News* (apud LAROSE, 2005), o que pode se justificar por ser uma tecnologia impar, passível de ser aplicada a uma variedade de domínios de problemas, além de ter como característica a multidisciplinaridade de seus algoritmos (que podem envolver áreas como estatística, matemática, inteligência artificial, recuperação de informação e processamento de sinais), o que faz com que melhorias possam surgir por meio de áreas e metodologias diversas.

2.3.1 Descoberta de Conhecimento em Bases de Dados

A aplicação de mineração de dados só obtém sucesso quando se pode garantir a integridade e acuracidade dos dados a serem usados. Para isso, existe um processo mais abrangente contendo fases que incluem desde a correta alimentação de bases de dados e definição de objetivos a serem alcançados no processo, até as fases de preparação, consolidação e, efetivamente, a mineração de dados (BERNARDI, 2010). Esse processo maior, do qual a mineração faz parte, é denominado descoberta de conhecimento em base de dados (do inglês: *Knowledge Discovery in Database* - KDD) (BRAGA, 2005). KDD, como

todo processo, consiste de uma seqüência iterativa de passos (HAN; KAMBER, 2006), que são retratados na Figura 2.9.

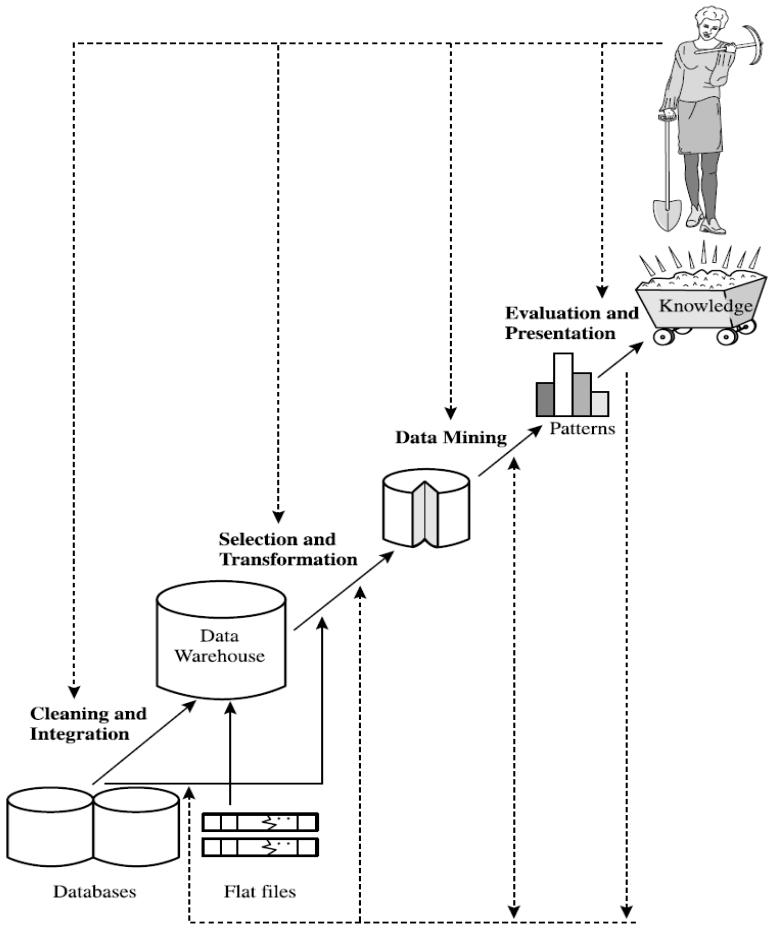


Figura 2.9. A mineração de dados como uma etapa do processo de KDD (HAN; KAMBER, 2006).

É na limpeza dos dados (primeiro passo) que dados corrompidos e inconsistentes são removidos. Essa etapa é seguida pela integração dos dados (segundo passo), quando dados de múltiplas fontes podem ser combinados. O terceiro passo corresponde à seleção de dados, onde as informações relevantes para a tarefa de análise são recuperadas do banco de dados. Na quarta etapa ocorre a transformação dos dados, quando os dados passam a ser representados de formas adequadas para a mineração. O quinto passo corresponde à mineração de dados, onde há a aplicação de algoritmos inteligentes com o objetivo de extrair padrões dos dados. Após a mineração, encontra-se a avaliação de padrões (sexto passo), quando são identificados os padrões verdadeiramente interessantes, que possibilitem a aquisição de novos conhecimentos. No sétimo e último passo, o conhecimento é apresentado, sendo usadas técnicas de visualização e representação de conhecimento para apresentar o conhecimento extraído para os usuários (HAN; KAMBER, 2006).

2.3.2 Modelo CRISP-DM

Segundo SHEARER (2000), em 1996, havia uma clara necessidade de um modelo de processo de mineração de dados que padronizasse a aplicação de mineração de dados e ajudasse as organizações a criar seus próprios projetos de mineração. A construção de um modelo não proprietário, documentado e livre permitiria às organizações obterem melhores resultados no processo de mineração e incentivaria o uso das melhores práticas na indústria, proporcionando a maturidade do mercado. Assim surgiu o *Cross-Industry Standard Process for Data Mining* (CRISP-DM), provendo um processo padrão para a aplicação apropriada de mineração de dados na resolução de problemas comerciais e de pesquisa (LAROSE, 2005).

O CRISP-DM estabelece um ciclo de vida dotado de seis fases para um projeto de mineração de dados. As fases obedecem a uma seqüência de execução adaptativa, onde as próximas fases a serem executadas dependem das saídas das fases anteriores. A Figura 2.10 ilustra o ciclo de vida segundo o CRISP-DM, onde as dependências mais comuns são indicadas pelas setas internas e o ciclo natural pelas externas (LAROSE, 2005).

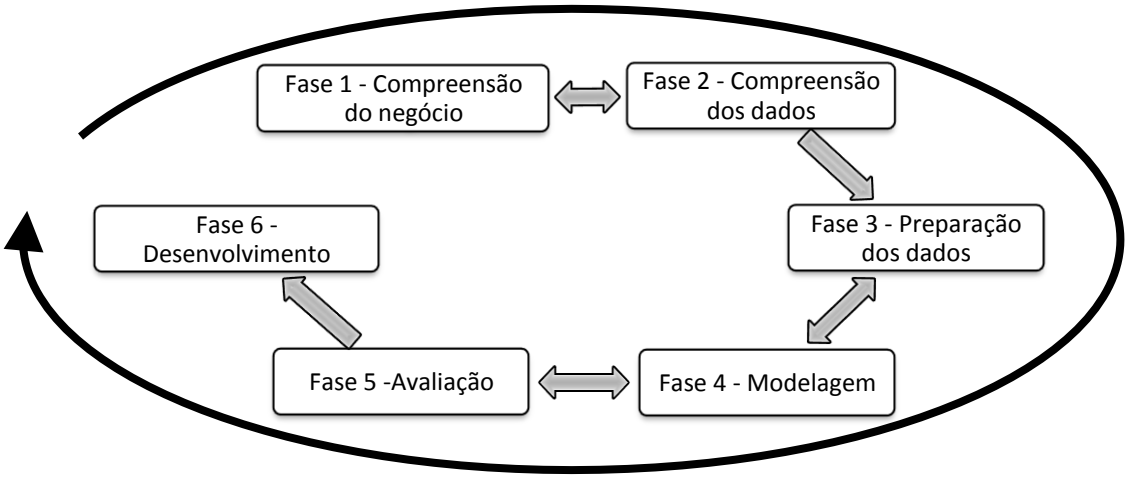


Figura 2.10. O processo iterativo e adaptável CRISP-DM (Adaptado de: LAROSE, 2005).

As seis fases do CRISP-DM, definidas por SHEARER (2000), são:

- Fase 1 – Compreensão do Negócio: Foca no entendimento dos objetivos do projeto a partir de uma perspectiva de negócios, convertendo esse conhecimento em uma definição do problema de mineração de dados, que é seguida pelo desenvolvimento de um plano preliminar para atingir os objetivos. Esta fase

compreende a determinação dos objetivos do negócio, a avaliação da situação, a determinação dos objetivos da mineração e a produção do plano do projeto.

- Fase 2 – Compreensão dos dados: Permite que, a partir de uma coleção inicial, o analista obtenha maior familiaridade com os dados, identificando problemas na qualidade, descobrindo ideias iniciais ou detectando subconjuntos interessantes para formar hipóteses sobre informações ocultas. Esta fase compreende a recolha dos dados iniciais, a descrição, a exploração e a verificação da qualidade dos dados.
- Fase 3 – Preparação dos dados: Abrange as atividades para a construção do conjunto final dos dados a partir dos dados brutos. As tarefas incluem: seleção, limpeza, construção, integração e formatação dos dados.
- Fase 4 – Modelagem: Onde as técnicas de modelagem são selecionadas e aplicadas, com devida calibração dos parâmetros a fim de obter bons resultados. Geralmente existem várias técnicas para o mesmo tipo de problema de mineração de dados, sendo que algumas possuem exigências específicas quanto a forma dos dados, o que pode necessitar um recuo para a fase de preparação. Esta fase compreende a seleção da técnica de modelagem, a geração de casos de teste, a criação e a avaliação de modelos.
- Fase 5 – Avaliação: Precede a implantação final do modelo construído pelo analista de dados. É fundamental para determinar se algum problema de negócio importante não foi suficientemente analisado, sendo que ao seu final deve-se decidir exatamente como usar os resultados da mineração de dados. Esta fase compreende a avaliação dos resultados, a revisão do processo e a determinação dos passos posteriores.
- Fase 6 - Desenvolvimento: Onde o modelo gerado anteriormente é usado e o conhecimento descoberto é apresentado ao interessado na mineração, para que este observe como pode usar o conhecimento obtido a seu favor. Esta fase compreende a implantação do plano, o monitoramento e manutenção do plano, a produção de um relatório final e a revisão do projeto.

2.3.3 Tarefas de Mineração de Dados

Segundo LAROSE (2005), a mineração de dados pode ser usada para realizar tarefas de descrição, estimativa, previsão, classificação, agrupamento e associação, sendo que existem diversos algoritmos destinados a cada tarefa. Estas tarefas encontrem-se descritas a seguir.

2.3.3.1 Descrição

A descrição é usada para encontrar formas de descrever padrões e tendências existentes em dados, como, por exemplo, para descobrir evidências de que funcionários que foram demitidos são menos propensos a apoiar o chefe em uma eleição presidencial. Descrições de padrões e tendências podem sugerir explicações para tais padrões e tendências, devendo o modelo de mineração de dados ser o mais transparente possível, a fim de que os resultados do modelo possam descrever padrões claros, que possam ser explicados e interpretados intuitivamente. Alguns métodos de mineração podem ser mais adequados para uma interpretação transparente, sendo que árvores de decisão podem fornecer uma explicação intuitiva e humanamente amigável dos seus resultados, enquanto que redes neurais podem apresentar-se de forma obscura para não especialistas, devido sua complexidade. A alta qualidade da descrição pode, muitas vezes, ser obtida pela análise exploratória de dados, um método gráfico de explorar dados na busca por padrões e tendências (LAROSE, 2005).

2.3.3.2 Estimativa

A estimativa constrói modelos com base em registros completos, através dos quais se estima o valor desconhecido de uma determinada variável em um novo registro. Um exemplo de aplicação seria estimar a nota média de um estudante de pós-graduação com base na nota média de graduação do aluno. Vários métodos de estimativa amplamente utilizados são fornecidos pelo campo de análise estatística, como estimativa pontual, estimativa de intervalos de confiança, regressão linear simples e múltipla (LAROSE, 2005).

2.3.3.3 Previsão

A previsão é similar a estimativa e a classificação, exceto que para a previsão os valores são previstos para o futuro. Um exemplo de aplicação seria a de prever o aumento de mortes no trânsito para o próximo ano se o limite de velocidade fosse aumentado. Os métodos e técnicas usados para a previsão podem ser os estatísticos tradicionais, como na estimativa,

além dos de descoberta de conhecimento, como redes neurais, árvores de decisão e dos vizinhos mais próximos (LAROSE, 2005).

2.3.3.4 Classificação

A classificação é similar a tarefa de estimativa, exceto que a variável alvo é categórica ao invés de numérica. A partir dessa variável, são especificadas classes discretas relacionadas aos registros da base de dados. Os dados já classificados são analisados, "ensinando" o método a classificar os outros registros, a partir das semelhanças. Um exemplo de aplicação seria identificar se um determinado comportamento pessoal ou financeiro indica uma possível ameaça terrorista. Métodos comuns de mineração usados para a classificação são os do vizinho mais próximo, árvores de decisão e redes neurais (LAROSE, 2005).

2.3.3.5 Agrupamento

O agrupamento consiste em agrupar registros em classes de objetos similares, sendo cada *cluster* uma coleção de registros que se assemelham entre si e que se diferem dos registros pertencentes aos outros *clusters*. A diferença do agrupamento para a classificação é que no agrupamento não há nenhuma variável alvo. Não se tenta classificar, estimar ou prever o valor de uma variável, mas sim segmentar os dados de um conjunto em subgrupos, nos quais a semelhança entre os registros internos é maximizada e a semelhança destes com os registros externos é minimizada. Um exemplo de aplicação seria para o agrupamento da expressão de genes, onde grandes quantidades de genes podem apresentar comportamento semelhante. (LAROSE, 2005).

2.3.3.6 Associação

A associação consiste na busca de relações entre dois ou mais atributos, prevalecendo no mundo dos negócios, onde é conhecida também como análise de afinidade. A associação pretende a descoberta de regras para quantificar a relação entre os atributos, sendo estas regras da forma "se antecedente, então conseqüente", juntamente com uma medida do apoio e confiança associada à regra. Um exemplo de aplicação seria para determinar a proporção de casos em que um novo medicamento apresentará efeitos secundários perigosos.

3 DESENVOLVIMENTO

Neste capítulo são descritas as atividades realizadas durante o desenvolvimento do trabalho. Na Seção 3.1 é apresentada a ferramenta de mineração de dados escolhida para uso, assim como comentários a cerca de sua instalação e configuração. Na Seção 3.2 é descrita a preparação dos dados para a mineração, incluindo a escolha dos dados relevantes para o problema, a obtenção e a formatação deles. Em seguida, na Seção 3.3, são apresentados os algoritmos implementados pela ferramenta de mineração que foram escolhidos para a realização dos testes.

3.1 Ambiente de Mineração WEKA

O WEKA foi escolhido por ser um *framework* livre de mineração de dados, consolidado nos ambientes acadêmico e científico. Além de ser multiplataforma, oferecer uma rica documentação e uma interface intuitiva, a ferramenta é bastante referenciada em artigos e livros de mineração de dados, como em BOUCKAERT et al. (2010), BRAGA (2005) e WITTEN et al. (2011).

O WEKA (*Waikato Environment for Knowledge Analysis*) é uma coleção de algoritmos de aprendizagem de máquina para tarefas de mineração de dados (University of Waikato, 2011). A ferramenta fornece também um conjunto de algoritmos de preparação de dados e de validação de resultados, tendo sido desenvolvido na Universidade de Waikato, na Nova Zelândia (SILVA, 2004). O sistema é escrito em Java, o que permite ser instalado em qualquer plataforma, e distribuído sobre os termos GNU *General Public License*, tendo recebido o nome de WEKA com o propósito de rimar com *Mecca*, uma ave encontrada apenas nas ilhas da Nova Zelândia (WITTEN et al., 2011).

3.1.1 Instalação e Configuração

A ferramenta WEKA pode ser obtida por meio de seu site (University of Waikato, 2011a), onde existe uma seção de download que disponibiliza instaladores para diferentes versões em diversas plataformas (Windows, Linux e MacOS). A instalação é bastante simples, bastando executar o instalador e dar seguimento às etapas, totalmente intuitivas, não sendo necessárias configurações adicionais para sua execução. Neste trabalho é usada a versão 3.6 para Windows 7 do ambiente, tendo sido necessário apenas incrementar a quantidade de memória utilizada pela máquina virtual Java antes da execução do WEKA, visto que nos primeiros testes o sistema acusava insuficiência de memória e encerrava automaticamente. Para aumentar a quantidade de memória, foi alterada a linha "maxheap=256m" do arquivo "RunWeka.ini", localizado na pasta de instalação da ferramenta, trocando o valor de 256 para 1024, aumentando o limite de memória as ser usada na execução do WEKA para 1 Gb.

3.1.2 Interface WEKA

O WEKA fornece uma interface uniforme para diferentes algoritmos de aprendizagem, pré e pós-processamento e avaliação, permitindo que seus usuários possam testar diversos métodos, identificando os mais apropriados para o problema em questão. As implementações de esquemas de aprendizagem reais podem ser consideradas o recurso mais valioso da ferramenta, sendo os recursos de pré-processamento de dados, que permite selecionar filtros em um menu e adequá-los diferentes necessidades, o segundo mais importante (WITTEN et al., 2011). Na Figura 3.1 é apresentada a tela inicial do WEKA, havendo quatro opções de acesso às funcionalidades: *Explorer*, *Experimenter*, *Knowledge Flow* e *Simple CLI*.

Figura 3.1. Tela inicial do ambiente WEKA, mostrando as opções de acesso.

- *Explorer*: É a interface mais popular e interativa do WEKA, sendo apresentada na Figura 3.2. Ela permite a exploração rápida de dados e suporta o carregamento e a filtragem destes, além da aplicação de diversos algoritmos de classificação, agrupamento, associação, seleção de atributos e visualização (BOUCKAERT et al., 2010). A *Explorer* foi usada neste trabalho para a aplicação dos algoritmos de mineração.

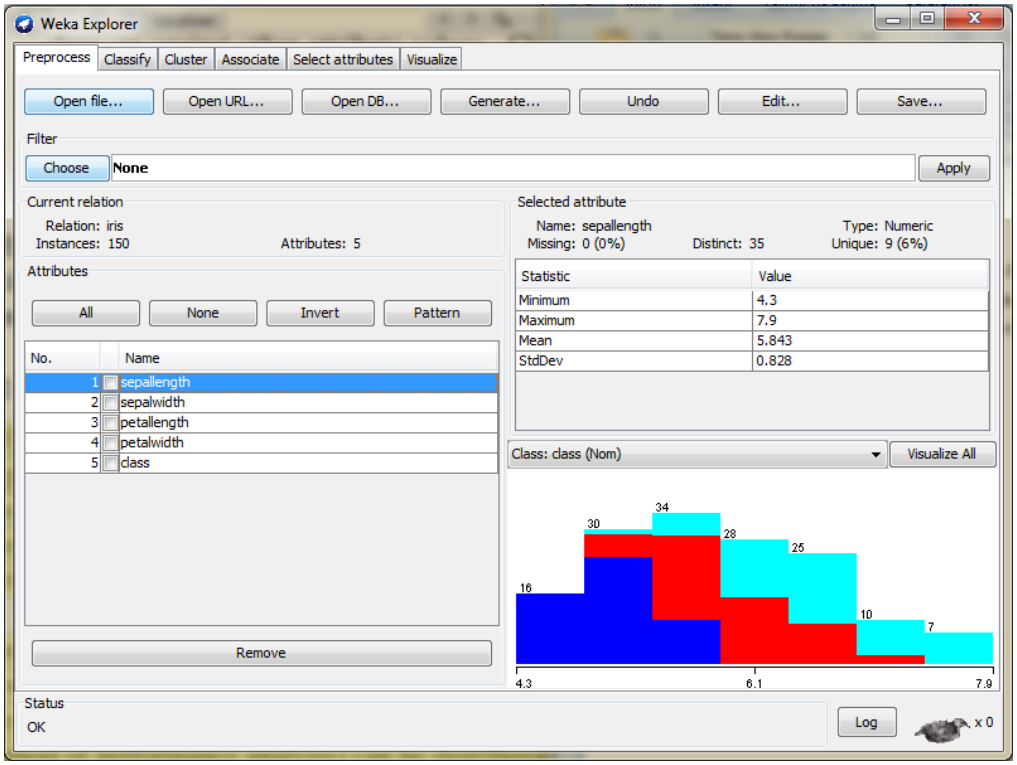


Figura 3.2. Interface *Explorer* do WEKA.

- *Experimenter*: É uma ferramenta para criação de experimentos de aprendizagem de máquina, que avalia métodos de classificação e de regressão. Ela permite comparação de desempenho e tabular resumos para incorporação em publicações. Nela, experimentos podem ser configurados para executar em paralelo sobre diferentes computadores em rede (BOUCKAERT et al., 2010). A *Experimenter* é representada na Figura 3.3 e foi usada neste trabalho para a escolha dos algoritmos de classificação usados.

- *Knowledge Flow*: Permite que configurações sejam projetadas para o processamento dos dados transmitidos. Pode-se especificar um fluxo de dados por componentes de conexão que podem representar fontes de dados, ferramentas de pré-processamento,

algoritmos de aprendizagem, métodos de avaliação e módulos de visualização (WITTEN et al., 2011).

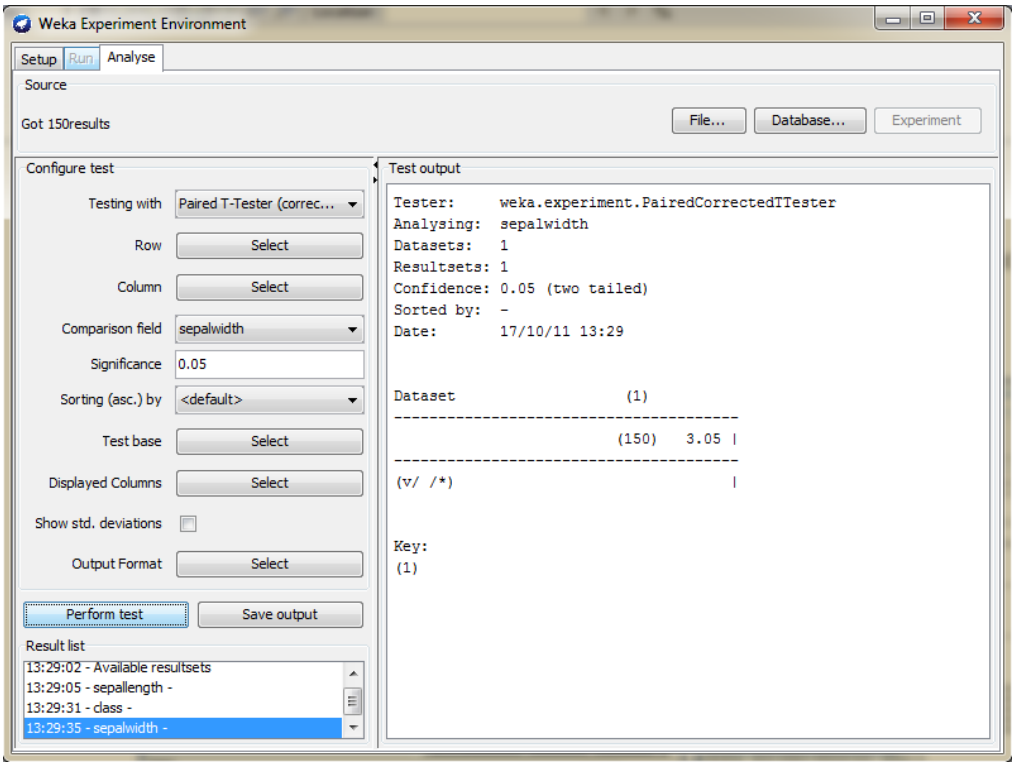


Figura 3.3. Interface *Experimenter* do WEKA.

- *Simple CLI*: Consiste no acesso às funcionalidades básicas do WEKA por meio de uma interface de linha de comando, que oferece um painel textual simples para a entrada de comandos, não sendo muito utilizado devido a oferta de formas mais intuitivas (WITTEN et al., 2011).

3.2 Preparação dos Dados

Com base no modelo CRISP-DM, em um processo de mineração de dados deve-se compreender e preparar os dados a serem utilizados antes da etapa de modelagem. As atividades relacionadas à preparação dos dados realizadas neste trabalho são apresentadas nesta seção.

3.2.1 Formato de Arquivo ARFF

O formato ARFF (*Attribute-Relation File Format*) é um arquivo de texto com codificação ASCII, que descreve uma lista de instâncias de um conjunto de atributos. Esse

formato foi desenvolvido pelo Projeto de Aprendizagem de Máquina no Departamento de Ciência da Computação da Universidade de Waikato para ser usado como formato padrão de entrada de dados do WEKA (University of Waikato, 2011b).

Os arquivos apresentam no início o cabeçalho das informações, que é seguido pelos dados propriamente ditos, como ilustrado na Figura 3.4. No cabeçalho, a declaração "@RELATION" descreve o nome da relação e a "@ATTRIBUTE" contém o nome do atributo seguido de seu tipo de dado. A seção de dados é iniciada com a declaração @DATA, seguida por linhas, que representam as instâncias, com valores de atributos separados por vírgulas e na ordem em que foram declarados no cabeçalho.

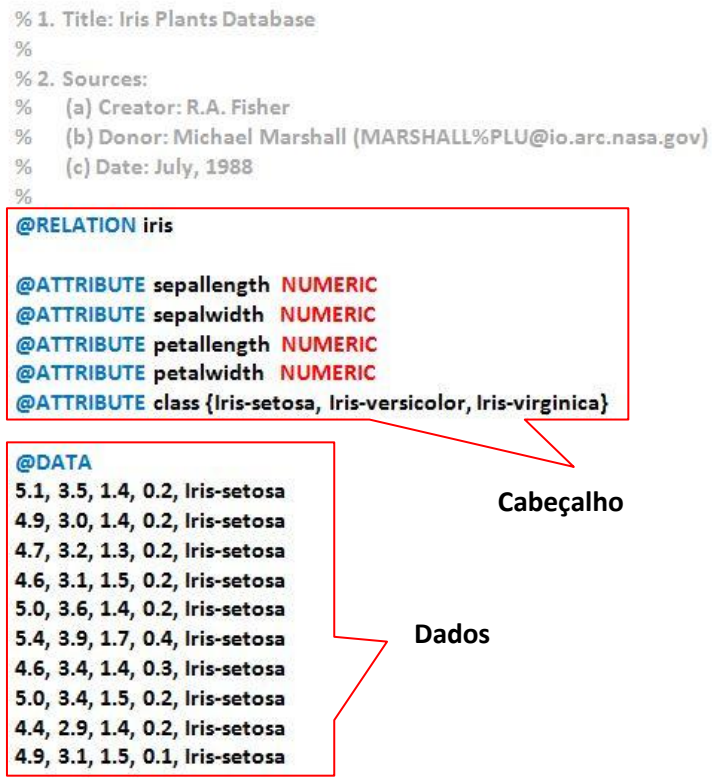


Figura 3.4. Seções de um arquivo ARFF.

3.2.2 Escolha de dados e estruturação de arquivos ARFF

Para o objetivo do trabalho, que é avaliar a relação entre tempestades geomagnéticas e dados de incidências de muons, o arquivo de dados de entrada para a mineração deve ser estruturado de modo a conter atributos que identifiquem a situação de tempestades e a contagem de muons para mesmos períodos temporais. Durante o desenvolvimento deste trabalho, foi criado e alterado várias vezes o arquivo de entrada de dados, modificando tanto a

formatação dos dados quanto os atributos usados, com a finalidade de melhorar a representatividade dos dados na posterior aplicação de algoritmos de mineração. As alterações mais significativas são apresentadas a seguir.

3.2.2.1 Uso de dados de ocorrência de CMEs e de variação de contagem de muons

A primeira etapa para construir o arquivo de dados de entrada foi pensar em dados que pudessem indicar a ocorrência de tempestades geomagnéticas para posteriormente buscar por suas fontes, visto que os dados de contagens de muons já podiam ser acessados através do banco de dados (MMD-DB) apresentado na Seção 2.2.1.3. Inicialmente, pensou-se que um bom indício de ocorrência de tempestades seria a ocorrência de CMEs, visto que as ejeções podem estar associadas à presença de tempestades. Na busca por fontes de dados de chegada de CMEs, foi encontrado o iSWA (*integrated Space Weather Analysis System*) (iSWA, 2011), um sistema desenvolvido pela NASA (*National Aeronautics and Space Administration*), que combina previsões baseadas nos mais avançados modelos de Clima Espacial com informações simultâneas do ambiente espacial. As previsões de tempo de chegada de CMEs são disponibilizadas pelo iSWA através de informações de saída, chegada e duração do distúrbio das CMEs, além de intervalos de confiança para as previsões, como mostrado na Figura 3.5.

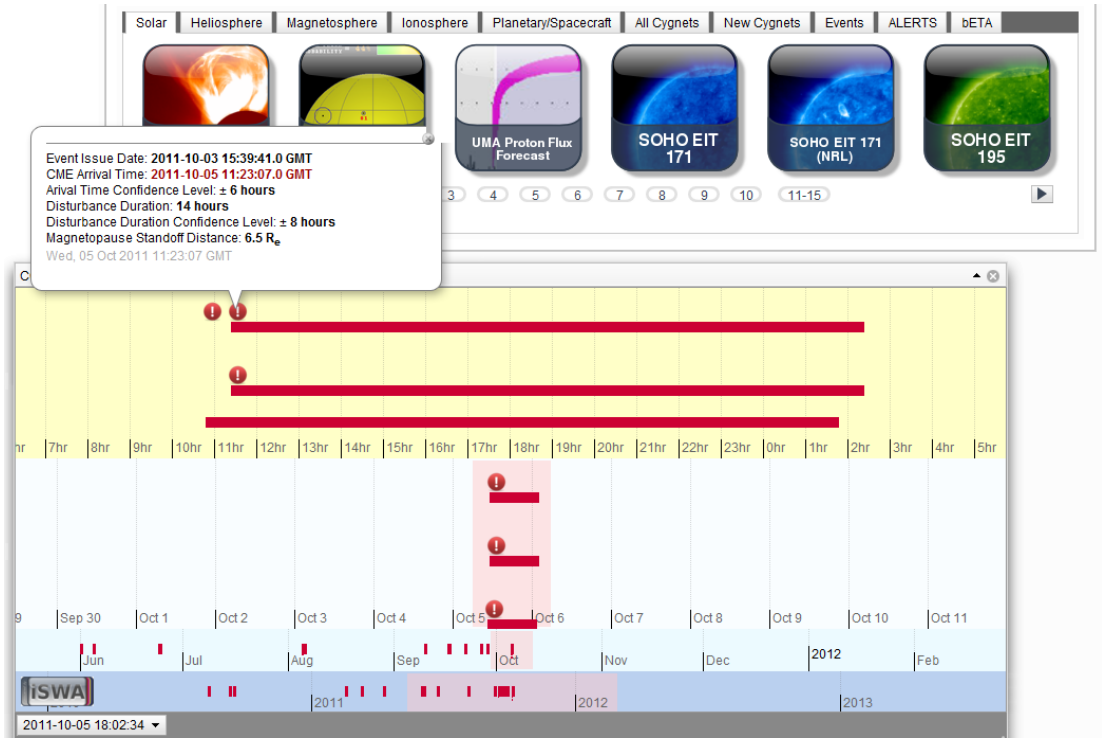


Figura 3.5. Previsão de tempo de chegada de CMEs a partir do sistema iSWA.

Com base nas informações do iSWA, foi feita uma busca por todos os dados de CMEs ocorridos a partir de dezembro 2006 (primeiro mês de dados de muons armazenados pelo MMD-DB) até junho de 2011, tendo encontrado dados de 14 CMEs, nenhum anterior a 2010, o que pode se justificar por este ser um período de baixa atividade solar.

Após as fontes definidas, foi feita a estruturação do arquivo, consistindo da definição dos atributos (e dos tipos de dados) que iriam participar do arquivo de entrada de dados. O fato de diversos fatores ambientais poderem influenciar as contagens de muons faz com que possa acontecer, por exemplo, que em uma determinada época do ano a incidência de muons seja reduzida, sem caracterizar necessariamente uma tempestade geomagnética. Dessa forma o valor absoluto de contagem para um determinado momento não é suficientemente representativo, sendo necessário um valor adicional que represente a variação da contagem, tendo sido usada para isso a diferença percentual da contagem para sua anterior (usando-se intervalos de uma hora). Para as CMEs, decidiu-se usar um atributo que especificasse a distância delas à Terra, sendo esse valor calculado de acordo com o tempo de saída e chegada da CME, considerando que antes da emissão da CME o valor da distância é igual a um e ao chegar a Terra o valor é igual a zero, sendo os valores dentro desse intervalo calculados proporcionalmente. A estrutura do arquivo em formato ARFF é apresentada na Figura 3.6.

```
@RELATION muons

% momento da contagem
@ATTRIBUTE date DATE "yyyy-MM-dd HH:mm:ss"

% soma das contagens de todas as direções
@ATTRIBUTE count NUMERIC

% diferença percentual do valor de count para o da instância anterior
@ATTRIBUTE diference NUMERIC

% distância da CME à superfície terrestre 1= CME não saiu     0 = CME chegou
@ATTRIBUTE distance NUMERIC

@DATA
"2010-01-01 00:00:00.0", 82060408, 0.0, 1
"2010-01-01 01:00:00.0", 81950532, -0.13389648269845308, 1
"2010-01-01 02:00:00.0", 82005002, 0.06646692665765733, 1
"2010-01-01 03:00:00.0", 82057475, 0.06398756017346356, 1
```

Figura 3.6. Arquivo ARFF para a abordagem baseada em dados de CMEs e variação de contagem de muons.

3.2.2.2 Uso de dados do índice Dst e de variação de contagem de muons

A pouca quantidade de dados de ocorrência de CMEs no arquivo de dados de entrada da Seção 3.2.2.1 restringiu a aplicação de algoritmos de mineração, visto que além de serem poucas as instâncias que caracterizavam CMEs, estas ainda deveriam ser divididas para treinamento e teste dos algoritmos.

Outro problema identificado, que desencorajou totalmente o uso dos dados de previsão de CMEs como indício de ocorrência de tempestades geomagnéticas, foi que para dados de algumas das CMEs, os dados de contagens de muons não apresentavam variação representativa, o que pode ser justificado pelo fato de estarem sendo usados dados de previsão. Na Figura 3.7 são mostrados graficamente os dados de contagem para a direção vertical, desde o momento de saída até o momento de chegada (com o retângulo vermelho marcando o intervalo de confiança estabelecido para a chegada) do único dado de CME do sistema iSWA para fevereiro de 2011, onde nota-se que as contagens apresentam um comportamento oscilatório padrão, sem decréscimos substanciais no intervalo de chegada da CME a Terra.

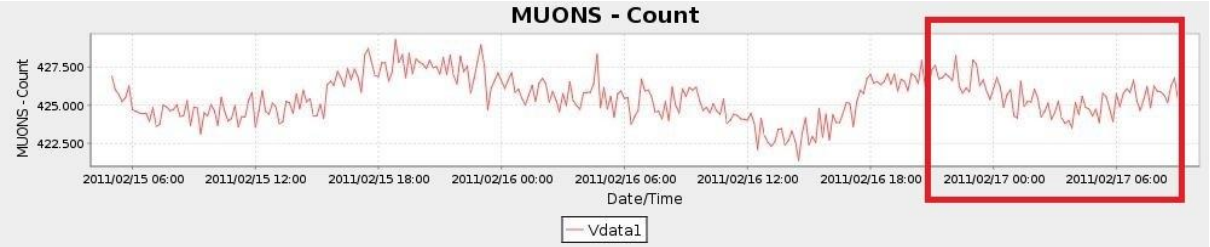


Figura 3.7. Contagens de muons na direção vertical para a CME de fevereiro de 2011 prevista pelo iSWA.

Com os problemas identificados no arquivo de dados inicialmente construído não teria como confiar nos resultados da aplicação de algoritmos de mineração, visto que além de poucos dados de ocorrência de CMEs, os dados inconsistentes também influenciariam o treinamento dos algoritmos. Dessa forma, buscou-se por outra informação que caracterizasse tempestades geomagnéticas para reestruturar e reconstruir o arquivo de entrada de dados, tendo sido encontrado o índice Dst, apresentado na Seção 2.1.3.1, que expressa o grau de perturbação do campo magnético.

Como fonte de dados do índice Dst, foi usado o *World Data Center* (WDC) *for Geomagnetism* da Universidade de Kyoto, no Japão, que disponibiliza um sistema de download de arquivos com dados de Dst, que pode ser encontrado em Kyoto University

(2011). Para gerar e fazer o download de um arquivo é necessário informar o período temporal desejado, marcar a opção “Dst Output” e o formato “IAGA2002” como formato de dados e informar um endereço de e-mail.

A nova estrutura do arquivo, mostrada em formato ARFF na Figura 3.8, substitui o atributo “distance” da anterior por um atributo nominal que define a situação de tempestades geomagnéticas a partir do valor de Dst (obtido a partir do arquivo de output gerado pelo WDC) de acordo com a Tabela 2.2 apresentada na Seção 2.1.3.1.

```
@RELATION muons

% momento da contagem
@ATTRIBUTE date DATE “yyyy-MM-dd HH:mm:ss”

% soma das contagens de todas as direções
@ATTRIBUTE count NUMERIC

% diferença percentual do valor de count para o da instância anterior
@ATTRIBUTE diference NUMERIC

% caracterização de tempestade, baseada no índice Dst
@ATTRIBUTE class {INEXISTENTE, FRACA, MODERADA, INTENSA, MUITO_INTENSA}

@DATA
"2010-01-01 00:00:00.0", 82060408, 0.0, INEXISTENTE
"2010-01-01 01:00:00.0", 81950532, -0.13389648269845308, INEXISTENTE
"2010-01-01 02:00:00.0", 82005002, 0.06646692665765733, INEXISTENTE
"2010-01-01 03:00:00.0", 82057475, 0.06398756017346356, INEXISTENTE
```

Figura 3.8. Arquivo ARFF para a abordagem baseada em dados de Dst e variação de contagem de muons.

3.2.2.3 Uso de dados do índice Dst e de média móvel para a contagem de muon

O arquivo de dados baseado no índice Dst forneceu uma quantidade maior de instâncias caracterizadas pela existência de tempestades geomagnéticas, o que agrega maior confiabilidade ao processo de mineração de dados. Porém, notou-se que o fato de as contagens de muons apresentarem um comportamento de contínua oscilação, faz com que, mesmo com as diferenças de uma contagem para sua anterior, a variação da contagem, comparada com o comportamento padrão, seja mascarada. Na Figura 3.9 pode ser observado um decaimento correspondente a uma tempestade intensa (segundo o índice Dst) em dezembro de 2006, e o apresenta em uma resolução temporal maior, para que seja visível a oscilação do valor das contagens. Embora se note um decaimento significativo através da análise do gráfico como um todo, analisando apenas a parte ampliada percebe-se que diferenças pontuais de contagens podem não ser capazes de indicar esse comportamento.

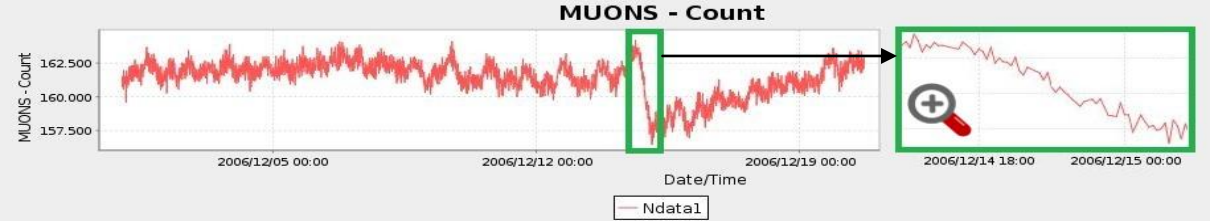


Figura 3.9. Comportamento oscilatório das contagens de muons.

A fim de minimizar esse problema, atenuando-se as oscilações de contagem, resolveu-se aplicar uma média móvel simples (com funcionamento ilustrado na Figura 3.10) ao somatório das contagens, sendo que o atributo "count" deixou de representar esse somatório e passou a receber o valor da média calculada. Dessa forma o atributo "diference" deixou de ser calculado a partir dos valores brutos das contagens e passou a ter seu valor dependente dos valores obtidos após a aplicação do filtro.

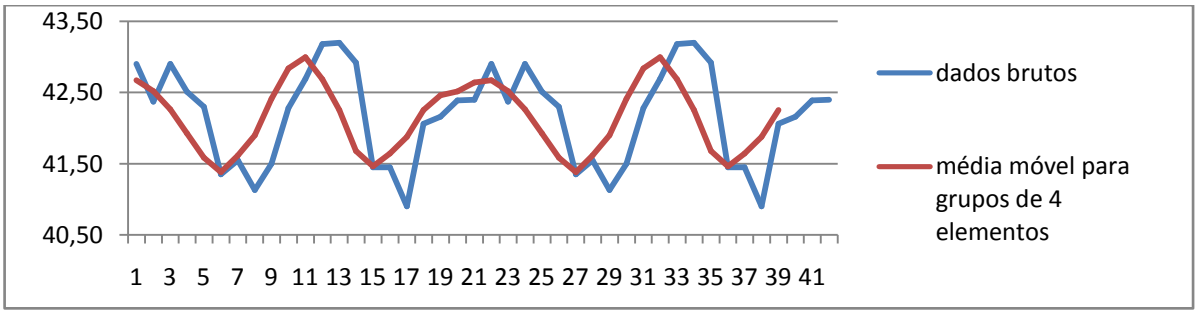


Figura 3.10. Funcionamento de uma média móvel.

3.2.3 Programa para gerar os arquivos ARFF

Para a criação dos arquivos de entrada de dados em formato ARFF foi desenvolvido um programa em Java que acessa os dados de contagens de muons do MMD-DB através do *framework* Hibernate (Hibernate, 2011), usado para realizar o mapeamento objeto relacional das tabelas do banco e facilitar o acesso e uso dos dados. Os dados de contagens de muons são associados aos de tempestades geomagnéticas de acordo com a equivalência dos atributos temporais de medição de ambas as informações. Inicialmente, os dados de CME eram inseridos via código, visto que eram conseguidos manualmente a partir do sistema iSWA, que apresentava as informações de previsão CMEs de forma visual. Com a mudança para uso de dados Dst, a entrada de dados referentes às tempestades geomagnéticas passou a ser feita através da passagem do endereço local do arquivo de output gerado pelo sistema do WDC.

3.2.3.1 Estrutura do Programa

O *software* é composto por arquivos responsáveis pela configuração da conexão com o banco de dados, pelo mapeamento objeto relacional das tabelas, pela representação dos dados de tempestades e pela criação dos arquivos ARFF, como é apresentado na Figura 3.11.

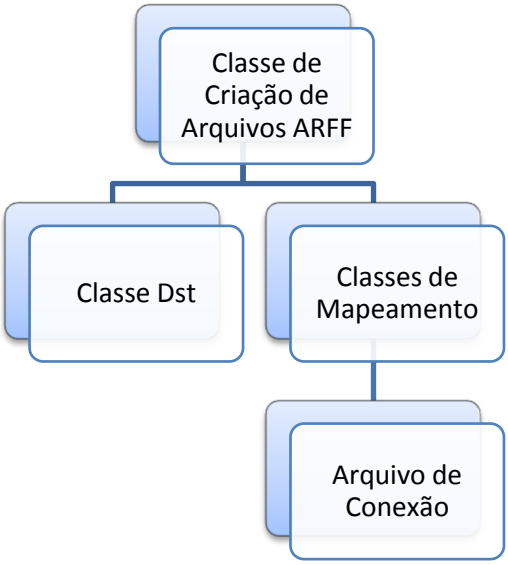


Figura 3.11. Arquivos que compõem o programa de criação de arquivos ARFF.

O arquivo "*hibernate.cfg.xml*" é responsável pela configuração do Hibernate para conexão com o banco de dados, onde são inseridas todas as informações necessárias para o acesso ao banco, como *driver*, usuário e senha do banco de dados a ser usado.

O mapeamento objeto relacional fez uso da conexão com o MMD-DB já estabelecida e foi realizado apenas para as tabelas necessárias para a aquisição dos dados de contagens em vários canais direcionais, resultando em quatro classes de mapeamento (diretamente associadas aos arquivos de dados de contagens de muon): "*MuonFile*", "*MuonData.java*", "*MuonDirectionalData* e "*Directions.java*".

Para a representação de dados de Dst que caracterizassem períodos de tempestades geomagnéticas, foi criada a classe "*Dst.java*", contendo atributos de horário de medição, valor do índice Dst medido e tipo da tempestade (baseado no valor do Dst). Essa classe conta com o método "*getDstList*" responsável por ler e interpretar o arquivo de dados de Dst, linha a linha, criando objetos de Dst referentes apenas a períodos caracterizados por tempestades e inserindo-os em uma lista, a qual é retornada.

A criação dos arquivos ARFF, incluindo o pareamento de dados de muons e de tempestades para a escrita das instâncias (cada uma em uma linha) no arquivo, é feita na classe "*DataForMining.java*", que é a classe principal do programa, onde se encontra o método "*main*'. Essa classe contém cinco métodos principais:

- "*getMuonsData*": responsável por retornar uma lista de objetos da classe de mapeamento "MuonData.java" para um determinado período de tempo.
- "*getTotalCount*": responsável por retornar o somatório total das contagens de muons em todos os canais direcionais para um determinado momento de medição.
- "*applyFilter*": responsável por retornar o valor resultante da aplicação de média móvel para os somatórios de contagens direcionais (obtidos pelo método "*getTotalCount*") a um determinado grupo de elementos.
- "*writeDataInFile*": responsável por escrever corretamente os dados passados por argumentos em uma linha no arquivo ARFF.
- "*main*": onde um novo arquivo contendo o cabeçalho dos dados ARFF é criado e os métodos "*getMuonsData*" e "*getDstList*" (da classe "Dst.java") são chamados, obtendo-se uma lista de objetos Dst referentes a períodos de tempestades e uma lista de dados de contagens de muons. A lista de contagens de muons é percorrida, sendo que para cada elemento da lista é chamado o método "*applyFilter*", para o cálculo da média móvel das contagens; buscado na lista de Dst o elemento com mesmo momento de medição, para realizar o pareamento dos dados; e chamado o método "*writeDataInFile*", para escrever os dados obtidos em uma linha do arquivo de dados.

3.3 Algoritmos de Mineração Aplicados

Dentre as tarefas de mineração de dados, a classificação foi considerada mais adequada para este trabalho, visto que se deseja treinar algoritmos de mineração para capacitá-los a classificar instâncias em uma variável categórica, que, no caso, representa a situação de tempestades geomagnéticas, que podem ser: inexistente, fraca, moderada, intensa ou muito intensa.

Nesta seção são apresentados os algoritmos de classificação escolhidos para serem aplicados ao arquivo de entrada ARFF. A escolha dos algoritmos se deu através do uso da interface "*experimenter*" do WEKA, onde comparações foram feitas, contatando-se que, para este caso, os algoritmos dos subpacotes *"trees*" e "*rules*" da aba "*classify*" do WEKA apresentaram melhores resultados. Dessa forma, foram escolhidos dois algoritmos de cada subpacote ("*J48*" e "*RandomTree*", para árvores de decisão; e "*DecisionTable*" e "*DTNB*", para regras de classificação), além de um algoritmo de meta-aprendizagem, "*AdaBoostM1*", usado para tornar um aprendiz "fraco" mais poderoso. Neste trabalho, o *AdaBoostM1* foi combinado ao algoritmo *DecisionTable*, que, dentre os quatro primeiros algoritmos escolhidos, foi o que apresentou o maior erro absoluto relativo, como será apresentado na Seção 4.1.3.

3.3.1 Árvores de Decisão

Árvores de decisão são baseadas na estratégia de dividir para conquistar, podendo ser expressas de forma recursiva, onde, primeiramente, um atributo é selecionado para ser colocado na raiz, tendo uma ramificação para cada possível valor e dividindo, assim, as instâncias de dados em subconjuntos. Esse processo é repetido de forma recursiva para cada ramo, usando apenas as instâncias que chegaram a ele, até que todas as instâncias em um nó tenham a mesma classificação, o que encerra o desenvolvimento da árvore nesse ramo. Na Figura 3.12 é apresentado um exemplo de árvore de decisão para a definição do tipo de lentes de contato que um paciente necessita, de acordo com os sintomas apresentados.

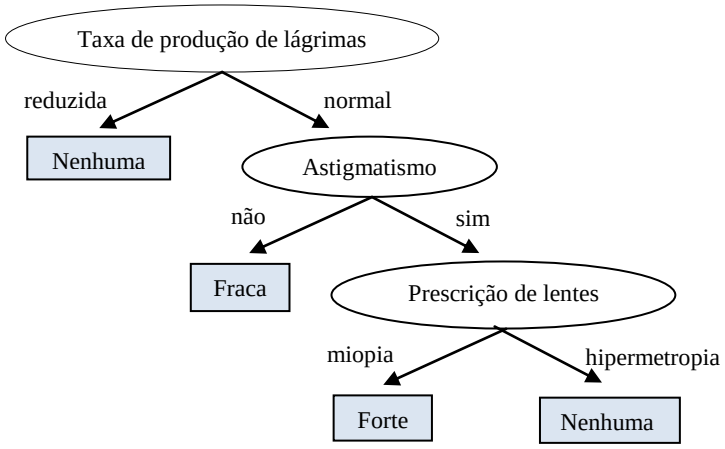


Figura 3.12. Árvore de decisão para dados de lentes de contato (Adaptado de: WITTEN et al., 2011).

Dessa forma, em uma árvore de decisão os nós envolvem testes para um atributo particular e as folhas representam classes, sendo que para classificar uma instância desconhecida, esta será encaminhada para baixo da árvore de acordo com os valores dos atributos testados em nós sucessivos até alcançar uma folha, que definirá sua classificação (WITTEN et al. 2011).

3.3.1.1 J48

O algoritmo *J48* é uma implementação em Java do algoritmo C4.5, para geração de árvores de decisão, e, segundo ALMEIDA (2003), é considerado o algoritmo mais popular do WEKA. O J48 se baseia em um conjunto de dados de treinamento para construir um modelo de árvore de decisão, que é usado para classificar as instâncias do conjunto de teste. Para a construção de uma árvore de decisão é necessário decidir como determinar o próximo atributo a ser usado como raiz da subárvore, sendo que o melhor atributo é aquele que resultará em uma árvore menor. Há várias heurísticas para essa escolha e no caso do *J48* (assim como na maioria dos algoritmos de árvore de decisão) a heurística é baseada no ganho de informação, sendo que esse ganho aumenta de acordo com a pureza[1] média dos subconjuntos que o atributo produz (WITTEN et al. 2011).

3.3.1.2 RandomTree

Uma árvore randômica é uma árvore construída aleatoriamente a partir de um conjunto de possíveis árvores, sendo que cada árvore do conjunto de árvores tem a mesma chance de ser amostrada, ou seja, a distribuição das árvores é uniforme. Árvores aleatórias clássicas são construídas a partir de um único vértice e, a cada iteração, um novo vértice, que tem uma probabilidade uniforme de se conectar a um vértice já existente, é adicionado. Segundo WITTEN et al. (2011), o algoritmo *RandomTree*, contido no WEKA, constrói uma árvore que considera um determinado número de características aleatórias em cada nó, sem executar nenhuma poda.

3.3.2 Regras de Classificação

Regras de classificação, também chamadas de regras de produção, constituem uma forma de representação simbólica, seguindo o formato: SE <antecedente> ENTÂO

[1] Um alto grau de pureza representa que uma classe é predominante em um conjunto de instâncias.

<consequente>. O antecedente refere-se a expressões condicionais envolvendo atributos dos dados de entrada e o consequente refere-se a uma expressão que indica um valor para um atributo meta, obtido em função dos valores dos atributos pertencentes ao antecedente. Algoritmos de regras de classificação geram um conjunto de regras a partir de um conjunto de dados de treinamento, sendo as regras resultantes usadas para classificar o restante das instâncias de dados, pertencentes a um conjunto de teste.

3.3.2.1 DecisionTable

O algoritmo *DecisionTable* constrói um classificador baseado na utilização de uma tabela de decisão com um mapeamento de regras para a classe majoritária. Essa representação é chamada de DTM (*Decision Table Majority*) e é constituída por um esquema, que é o conjunto de características que estão incluídas na tabela, e um corpo, consistido de instâncias rotuladas no espaço definido pelas características definidas no esquema. Para classificar uma instância, o classificador procura por casamentos perfeitos dentro da tabela, usando apenas as características do esquema, e retorna a classe majoritária da DTM, caso nenhuma instância seja encontrada, ou as classes majoritárias de todas as instâncias encontradas (KOHAVI, 1995). Uma opção do algoritmo usa o método de vizinho mais próximo para determinar a classe de cada instância não encontrada na tabela de decisão, com base no mesmo conjunto de características, ao invés da maioria global da tabela (WITTEN et al. 2011).

3.3.2.2 DTNB

DTNB é um classificador híbrido que combina uma tabela de decisão com o classificador Naive Bayes, um classificador simples e intuitivo que assume que os atributos são condicionalmente independentes (por isso denominado ingênuo) e que se baseia na regra de Bayes de probabilidade condicional. O *DTNB* divide os atributos em dois grupos, um modelado pela tabela e outro por Naive Bayes. Inicialmente, todos os atributos são modelados pela tabela de decisão, sendo feita uma busca gananciosa para decidir quais atributos devem ser modelados por Naive Bayes. Ao final, as previsões geradas pelos dois métodos são combinadas, usando a regra de Bayes (WITTEN et al. 2011).

3.3.3 Meta-Aprendizagem

Algoritmos de meta-aprendizagem têm como parâmetro um classificador base, e possuem o objetivo de transformar esse classificador em um aprendiz mais poderoso, podendo

seguir diversas abordagens para isso, como a de *Boosting*, mencionada a seguir, na Subseção 3.3.3.1.

3.3.3.1 AdaBoostM1

Boosting é um método geral para melhorar o desempenho de algoritmos de aprendizagem considerados fracos (embora também possa ser combinado a algoritmos fortes, como o C4.5). Esse método executa repetitivamente um determinado algoritmo de aprendizagem em várias distribuições sobre os dados de treinamento, combinando, ao final, os classificadores produzidos pelo algoritmo de aprendizagem em um único classificador composto (FREUND, 1996). O método de *Boosting* deriva os modelos individuais através de uma ponderação, que é usada para dar mais influência aos mais bem sucedidos. O algoritmo *AdaBoostM1*, contido no WEKA, implementa o método de *Boosting* e pode ser acelerado através da especificação de um limite para a poda de pesos (WITTEN et al. 2011).

4 RESULTADOS

Neste capítulo são apresentados os resultados obtidos através da aplicação dos algoritmos de mineração escolhidos (explicados na Seção 3.3) a um arquivo ARFF de entrada de dados, gerado pelo programa cuja implementação foi detalhada na Seção 3.2.3. O arquivo usado contem dados que vão desde 01 de dezembro de 2006 até 30 de outubro de 2011, tendo sido usada uma média móvel simples com grupos de três elementos para a atenuação das oscilações de contagens.

A aplicação de cada um dos algoritmos escolhidos foi feita por meio da interface *Explorer* do WEKA, utilizando os valores padrões de parâmetros e k-validação cruzada, com 10 subconjuntos. A validação cruzada divide a amostra em ‘k’ grupos (ou *folds*) de mesmo tamanho, sendo que o classificador é treinado com ‘k-1’ desses subconjuntos e testado para o *fold* restante. Este procedimento é repetido por ‘k’ vezes, cada uma usando um subconjunto de validação diferente, sendo que, ao final, a taxa de acerto é uma média das taxas de acerto nas ‘k’ iterações realizadas, fazendo desse método uma boa opção para medir o desempenho dos algoritmos e dos erros.

O desempenho dos algoritmos é apresentado na saída de resultados do WEKA por meio de diversas medidas e estatísticas, organizadas em seções separadas (*Summary*, *Detailed Accuracy By Class* e *Confusion Matrix*). A fim de facilitar a compreensão dos resultados e uma possível comparação entre os algoritmos, as principais medidas de desempenho são explicadas, de acordo com WITTEN et al. (2011), separadamente nesta seção e acompanhadas de seus resultados para cada um dos algoritmos testados. O agrupamento das medidas, respectivo às Seções 4.1, 4.2 e 4.3, segue a mesma divisão apresentada nos resultados do WEKA, mencionada anteriormente.

4.1 Resumo

Nesta seção são apresentados valores estatísticos que resumem o modo como os classificadores estão habilitados a predizer as classes corretas.

4.1.1 Instâncias Classificadas Correta e Incorretamente

Indicam o total e a porcentagem de classificações corretas e incorretas, respectivamente, realizadas por um algoritmo. Nas Tabelas 4.1 e 4.2, são apresentados esses valores para todos os algoritmos aplicados.

Tabela 4.1. Resultados obtidos pelos algoritmos para instâncias classificadas corretamente.

	DecisionTable	***DTNB***	***J48***	***RandomTree***	***AdaBoostM1***
Nº de Instâncias	40864	40859	40942	40972	41117
Porcentagem	97.7304 %	97.7184 %	97.9169 %	97.9887 %	98.3354 %

Tabela 4.2. Resultados obtidos pelos algoritmos para instâncias classificadas incorretamente.

	DecisionTable	***DTNB***	***J48***	***RandomTree***	***AdaBoostM1***
Nº de Instâncias	949	954	871	841	696
Porcentagem	2.2696 %	2.2816 %	2.0831 %	2.0113 %	1.6646 %

Analisando as resultados, percebe-se que, para todos os algoritmos, as porcentagens relativas às classificações corretas foram bastante altas (e, consequentemente, as de classificações incorretas bastante baixas), tendo sido o algoritmo *DTNB*, de regras de classificação, o que obteve menor taxa (embora também tenha sido alta). Outro fato a ser observado é a porcentagem (a maior) obtida pelo algoritmo *AdaBoostM1*, que usou como classificador base o *DecisonTable*, o qual teve, realmente, seu poder de aprendizagem aumentado. Embora os valores tenham sido todos altos, com essas informações, não se tem como verificar que todas as classes (inexistente, fraca, moderada, intensa, muito intensa), referentes a situação das tempestades geomagnéticas, tiveram taxas semelhantes de acertos.

4.1.2 Estatística *Kappa*

Índice que compara o valor encontrado nas observações com o valor que se pode esperar do acaso. Ele varia de 0 a 1 e quanto menor seu valor, menor a confiança da observação. Na Tabela 4.3, são apresentados os valores obtidos para esse índice pelos algoritmos testados.

Tabela 4.3. Resultados obtidos pelos algoritmos para o índice *Kappa*.

	DecisionTable	***DTNB***	***J48***	***RandomTree***	***AdaBoostM1***
Estatística *Kappa*	0.703	0.7018	0.7302	0.7574	0.8002

Pode-se notar que a ordem de desempenho dos algoritmos, tanto para a classificação de instâncias de forma correta, quanto para o índice *Kappa*, manteve-se a mesma, ou seja, os algoritmos que classificaram um número maior de instâncias corretamente, segundo a Seção 4.1.2, apresentaram maior confiabilidade de acordo com o índice *Kappa*.

4.1.3 Erro Absoluto Relativo

É o erro absoluto total calculado em relação ao que o erro deveria ter sido se a previsão fosse a média dos valores reais. Valores mais baixos de erro significam maior precisão do modelo e, dessa forma, um valor próximo de zero equivale a um modelo estatisticamente perfeito. Na Tabela 4.4, os erros absolutos relativos obtidos pelos algoritmos aplicados são apresentados.

Tabela 4.4. Resultados obtidos pelos algoritmos para o erro absoluto relativo.

	DecisionTable	***DTNB***	***J48***	***RandomTree***	***AdaBoostM1***
Erro absoluto relativo	51.9196 %	51.3286 %	34.9907 %	24.0813 %	19.9294 %

Nota-se que os valores de erro foram, em geral, menores para os algoritmos de maior confiabilidade (maior índice *Kappa*), sendo que os algoritmos de árvore de decisão, mais uma vez, obtiveram melhores resultados que os de regras de classificação e o algoritmo *AdaBoostM1*, de meta-aprendizagem, também obteve uma redução de erro significativa, comparando-o com o classificador base (*DecisionTable*). Pode-se observar também que,

mesmo sendo a diferença pequena, o algoritmo *DecisionTable* obteve uma taxa de erro maior que o *DTNB*, que possui maior índice *Kappa*.

4.2 Acurácia Detalhada por Classe

Nesta seção são apresentadas medidas referentes à acurácia da predição do classificador, detalhadas para cada uma das classes (inexistente, fraca, moderada, intensa, muito intensa) referentes ao atributo de situação das tempestades geomagnéticas.

4.2.1 *F-Measure*

Medida usada para mensurar o desempenho de um classificador, combinando valores de duas outras medidas, revocação e precisão, em uma única fórmula:

$$\frac{2 \times \text{revocação} \times \text{precisão}}{\text{revocação} + \text{precisão}}$$

Precisão é o valor da predição positiva, sendo o resultado da divisão do número de casos positivos pelo total de casos cobertos, enquanto revocação é o valor da cobertura de casos, tendo seu cálculo feito através da divisão do número de casos cobertos pelo número total de casos aplicáveis.

Na Tabela 4.5 são apresentados os valores para *F-Measures* obtidos por cada classe individualmente, para cada um dos algoritmos aplicados.

Tabela 4.5. Resultados obtidos para *F-Measure*, por classe.

	DecisionTable	***DTNB***	***J48***	***RandomTree***	***AdaBoostM1***
Inexistente	0.99	0.99	0.991	0.991	0.993
Fraca	0.663	0.662	0.69	0.724	0.769
Moderada	0.729	0.727	0.755	0.778	0.807
Intensa	0.87	0.87	0.784	0.857	0.923
Muito Intensa	-	-	-	-	-

Analisando os resultados da Tabela 4.5, percebe-se que, para todos os algoritmos, a classe que obteve valores maiores, indicando melhor desempenho, foi a “Inexistente”, seguida

por “Intensa”, “Moderada” e “Fraca”, nesta ordem. A classe “Muito Intensa” não foi avaliada por não existir nenhuma instância referente a uma tempestade deste tipo no arquivo de dados de entrada usado. O fato das classes extremas (“Inexistente” e “Intensa”) apresentarem melhores resultados talvez possa ser justificado por elas possuírem características mais expressivas, enquanto a classe “Fraca”, por exemplo, é muitas vezes confundida com a “Inexistente”, por ser caracterizada por valores de atributos não muito distantes desta.

4.2.2 Área ROC

A Curva ROC (*Receiver Operating Characteristic*) descreve o desempenho de um classificador desconsiderando a distribuição de classes ou custos de erros. Ela plota a taxa de verdadeiros positivos (número de positivos incluídos na amostra) no eixo vertical e a taxa de verdadeiros negativos (número de negativos incluídos na amostra) no eixo horizontal. A área ROC indica a probabilidade que uma instância positiva escolhida aleatoriamente no conjunto de dados de testes tem de ser classificada acima de uma instância negativa escolhida aleatoriamente, com base no *ranking* produzido pelo classificador. O melhor resultado é quando todas as instâncias positivas são classificadas acima de todas as negativas, gerando um valor de área ROC igual a um.

Na Tabela 4.6, são apresentados os valores de área ROC obtidos por cada classe individualmente, para cada um dos algoritmos aplicados.

Tabela 4.6. Resultados obtidos para a área ROC, por classe.

	DecisionTable	***DTNB***	***J48***	***RandomTree***	***AdaBoostM1***
Inexistente	0.932	0.932	0.917	0.892	0.963
Fraca	0.928	0.928	0.886	0.855	0.953
Moderada	0.957	0.961	0.933	0.884	0.982
Intensa	0.999	0.999	0.94	0.92	1
Muito_Intensa	-	-	-	-	-

Analisando-se a Tabela 4.6, pode-se notar que a classe com melhor desempenho, segundo a medida de área ROC, para todos os algoritmos foi a “Intensa”, tendo inclusive alcançado o valor máximo (um) para o “*AdaBoostM1*”, o que significa que, neste caso, todas as instâncias positivas foram classificadas acima das negativas no gráfico da curva ROC.

Outras observações a serem feitas são as de que, para essa medida, os algoritmos de regras de associação obtiveram maiores (e melhores) valores do que os de árvores de decisão e, embora, em geral, as classes “Fraca” e “Moderada” tenham obtido os menores valores, todos os valores (de todas as classes e algoritmos) foram bastante altos, o que caracteriza um bom desempenho.

4.3 Matriz de Confusão

A matriz de confusão apresenta a forma como as instâncias foram classificadas entre as classes, mostrando o número de classificações corretas para cada classe em oposição às classificações preditas erroneamente para outras. Ela é uma matriz quadrada nxn, onde ‘n’ é o número de classes, sendo que os verdadeiros positivos encontram-se na diagonal principal e os falsos positivos no restante da matriz. Dessa forma, se uma classificação for 100% correta, a matriz de confusão deverá ser uma matriz diagonal. Na Figura 4.1, é apresentado um exemplo de matriz de confusão para o conjunto de dados clássico da flor Íris. Neste exemplo, a Íris Setosa obteve 49 instâncias classificadas corretamente (verdadeiros positivos) e apenas uma erroneamente (falso positivo), que foi classificada como Íris Versicolor.

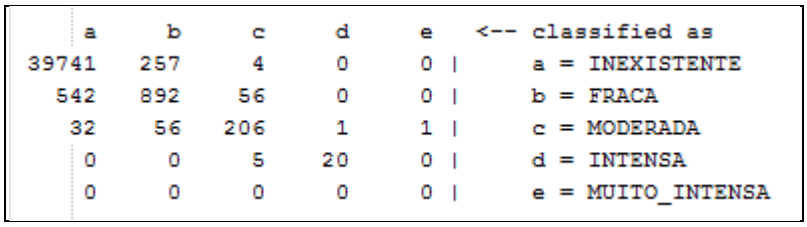
```
=== Confusion Matrix ===

  a  b  c   <-- classified as
 49  1  0 |  a = Iris-setosa
  0 47  3 |  b = Iris-versicolor
  0  2 48 |  c = Iris-virginica
```

Figura 4.1. Exemplo de uma matriz de confusão apresentada pelo WEKA.

A seguir, nas Figuras 4.2, 4.3, 4.4, 4.5 e 4.6, são apresentadas as matrizes de confusão obtidas para os algoritmos usados.

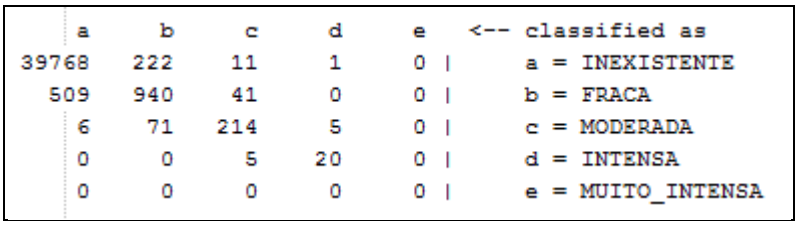
```
    a     b     c     d     e   <-- classified as
39745   253     4     0     0 |     a = INEXISTENTE
  542   892    56     0     0 |     b = FRACA
   32    56   207     1     0 |     c = MODERADA
    0     0     5    20     0 |     d = INTENSA
    0     0     0     0     0 |     e = MUITO_INTENSA
```

Figura 4.2. Matriz de confusão para o algoritmo *DecisionTable*.



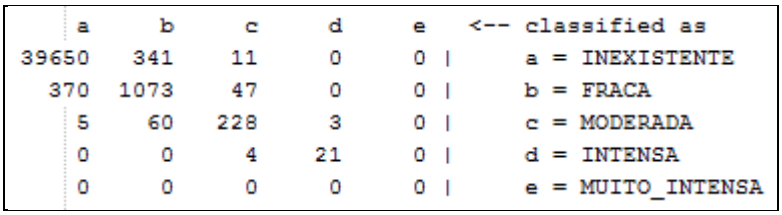
```
    a     b     c     d     e   <-- classified as
39741   257     4     0     0 |     a = INEXISTENTE
  542   892    56     0     0 |     b = FRACA
   32    56   206     1     1 |     c = MODERADA
    0     0     5    20     0 |     d = INTENSA
    0     0     0     0     0 |     e = MUITO_INTENSA
```

Figura 4.3. Matriz de confusão para o algoritmo *DTNB*.



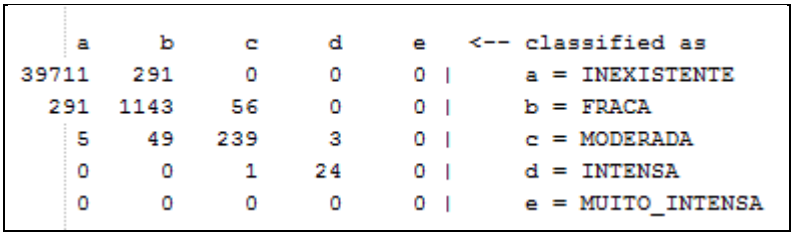
```
    a     b     c     d     e   <-- classified as
39768   222    11     1     0 |     a = INEXISTENTE
  509   940    41     0     0 |     b = FRACA
    6    71   214     5     0 |     c = MODERADA
    0     0     5    20     0 |     d = INTENSA
    0     0     0     0     0 |     e = MUITO_INTENSA
```

Figura 4.4. Matriz de confusão para o algoritmo *J48*.



```
    a     b     c     d     e   <-- classified as
39650   341    11     0     0 |     a = INEXISTENTE
  370  1073    47     0     0 |     b = FRACA
    5    60   228     3     0 |     c = MODERADA
    0     0     4    21     0 |     d = INTENSA
    0     0     0     0     0 |     e = MUITO_INTENSA
```

Figura 4.5. Matriz de confusão para o algoritmo *RandomTree*.



```
    a     b     c     d     e   <-- classified as
39711   291     0     0     0 |     a = INEXISTENTE
  291  1143    56     0     0 |     b = FRACA
    5    49   239     3     0 |     c = MODERADA
    0     0     1    24     0 |     d = INTENSA
    0     0     0     0     0 |     e = MUITO_INTENSA
```

Figura 4.6. Matriz de confusão para o algoritmo AdaBoostM1.

Todas as matrizes de confusão apresentadas classificaram um número muito maior de instâncias referentes à inexistência de tempestades corretamente (o que pode ser explicado pelo fato do arquivo de dados possuir muito mais instâncias desse tipo, havendo mais dados para treinamento e teste) e apresentaram maior confusão para as classes “Fraca” e “Moderada”, o que já era esperado, de acordo com os resultados da Seção 4.2.

Outra observação que pode ser feita é a de que, para todos os algoritmos, a classe “Fraca” foi mais confundida com a “Inexistente” durante a classificação das instâncias, da mesma forma que a “Moderada” teve instâncias classificadas incorretamente em maior número para a classe “Fraca”, possivelmente devido à proximidade dos valores de atributos

que as caracterizam. Além disso, os algoritmos de árvores de decisão, *J48* e *RandomTree* (Figuras 4.4 e 4.5), distribuíram mais corretamente as instâncias entre as classes do que os de regras de classificação, *DecisionTable* e *DTNB* (Figuras 4.2 e 4.3), e o algoritmo de meta-aprendizagem, *AdaBoostM1* (Figura 4.6), usado para tornar o classificador base *DecisionTable* (Figura 4.2) um aprendiz mais poderoso, confirmou seu bom desempenho, o que já era esperado, de acordo com os resultados vistos na Seção 4.1.

5 CONCLUSÃO

Neste trabalho foi apresentado todo o processo de aplicação de mineração a dados de incidência de muons, oriundos do detector brasileiro, e de índice Dst, indicador de perturbações no campo geomagnético, com o apoio do ambiente de mineração de dados WEKA, a fim de avaliar a associatividade entre essas partículas secundárias (muons) e a ocorrência de tempestades geomagnéticas. Esse processo abrangeu etapas como a preparação dos dados (incluindo seleção, integração e transformação destes) e o desenvolvimento de um software responsável por armazenar os dados escolhidos em um arquivo ARFF (formato padrão de entrada de dados do WEKA), além da escolha e a aplicação de algoritmos de mineração.

Para o domínio trabalhado, escolheu-se a técnica de classificação, sendo que os algoritmos de mineração aplicados foram escolhidos por meio de comparações e testes executados com o apoio da interface *Experimenter* do WEKA. Foram escolhidos dois algoritmos de classificação baseados em árvores de decisão, *J48* e *RandomTree*, dois em regras de classificação, *DecisionTable* e *DTNB*, e um de meta-aprendizagem, *AdaBoostM1*, que foi combinado ao algoritmo *DecisionTable* (que apresentou a maior taxa de erro relativo absoluto dentre os quatro primeiros algoritmos) a fim de torná-lo um aprendiz mais poderoso.

A análise dos resultados obtidos pela execução dos algoritmos de mineração, sobre o arquivo ARFF de dados trabalhado, confirma a relação entre a incidência de muons e a ocorrência de tempestades geomagnéticas, visto que, embora a classificação das instâncias em determinadas classes tenha apresentado uma taxa de acertos menor, todas as classes tiveram a maioria de suas instâncias classificadas corretamente. Isto significa que os algoritmos detectaram e aprenderam determinados padrões dos conjuntos de dados de treinamento, que

os capacitaram a classificar corretamente uma grande quantidade de instâncias dos conjuntos de dados de teste.

Como trabalhos futuros, poderiam ser usados dados da incidência de muons medidos por outros detectores, que não o brasileiro, ou até mesmo outros indicadores de tempestades geomagnéticas, a fim de validar os resultados obtidos através deste trabalho.

REFERÊNCIAS

ALMEIDA, L. M.; PADILHA, T. P. P.; OLIVEIRA, F. L; PREVIERO, C. A. Uma Ferramenta para Extração de Padrões. **Revista Eletrônica de Iniciação Científica**, v. 3, n.5, 2003.

BERNARDI, E. F. F. **Uma Arquitetura para Suporte à Mineração de Dados Paralela e Distribuída em Ambientes de Computação de Alto Desempenho**. 2010. Dissertação (Mestrado) - Pontífica Universidade Católica do Rio Grande do Sul.

BOUCKAERT, R. R.; FRANK, E.; HALL, M. A.; HOLMES, G.; PFAHRINGER, B.; REUTEMANN, P.; WITTEN, I. H. WEKA - Experiences with a Java Open-Source Project. **Journal of Machine Learning Research**, v. 11, p. 2533-2541, 2010.

BRAGA, L. P. V. **Introdução à Mineração de Dados**. 2ª. ed. Rio de Janeiro: E-Papers, 2005.

DAL LAGO, A. **Estudo de Estruturas Geoefetivas no Meio Interplanetário e de suas Causas Solares**. 2003. Tese (Doutorado em Geofísica Espacial) - Instituto Nacional de Pesquisas Espaciais.

DAL POZ, W. R.; CAMARGO, P. O. Consequências de uma Tempestade Geomagnética no Posicionamento Relativo com Receptores GPS de Simples Freqüência. **Boletim de Ciências Geodésicas**. Curitiba, v. 12, n. 2, p.275-294, 2006.

ECHER, E.; RIGOZO, N. R.; NORDEMANN, D. J. R.; VIEIRA, L. E. A.; PRESTES, A.; FARIA, H. H. O Número de Manchas Solares, Índice da Atividade do Sol. **Revista Brasileira de Ensino de Física**, v. 25, n. 2, p. 157-163, 2003.

ELMASRI, R.; NAVATHE, S. B. **Sistemas de Banco de Dados**. 4ª. ed. São Paulo: Pearson-Addison-Wesley, 2005.

EMBRACE – Estudo e Monitoramento Brasileiro do Clima Espacial. **Introdução ao Clima Espacial**. Disponível em: http://www.inpe.br/climaespacial/introducao.php. Acesso em: Outubro de 2011.

FREUND, Y.; SCHAPIRE, R. E. Experiments with a new boosting algorithm. In: **International Conference on Machine Learning**. Bari, p.148-156, 1996.

HAN, J.; KAMBER, M. **Data Mining: Concepts and Techniques**. 2ª. ed. San Francisco: Morgan Kaufmann, 2006.

HAND, D.; MANNILA, H.; SMYTH, P. **Principles of Data Mining**. 1ª. ed. Cambridge: MIT Press, 2001.

Hibernate. **Relational Persistence for Java and .NET**. Disponível em: http://www.hibernate.org/. Acesso em: Outubro de 2011.

iSWA - integrated Space Weather Analysis System. **The integrated Space Weather Analysis System**. Disponível em: http://iswa.gsfc.nasa.gov:8080/IswaSystemWebApp. Acesso em: Outubro de 2011.

Kyoto University. **Plot and data output of Dst and AE indices (Hourly Values)**. Disponível em: http://wdc.kugi.kyoto-u.ac.jp/dstae/index.html. Acesso em: Outubro de 2011.

KOHAVI, R. The Power of Decision Tables. In: **VIII European Conference on Machine Learning**. Heraklion, p.174-189, 1995.

KORTH, H. F.; SILBERSCHATZ, A.; SUDARSHAN, S. **Sistemas de Banco de Dados**. 5ª. ed. Rio de Janeiro: Campus, 2006.

LAROSE, D. T. **Discovering Knowledge in Data:** an introduction to data mining. 1ª. ed. Hoboken: Wiley-Interscience, 2005.

MATSUOKA, M. T.; COLLISCHONN, C.; CAMARGO, P. O. Análise do desempenho do Modelo Global da Ionosféra do IGS no posicionamento por ponto durante períodos de tempestades geomagnéticas: estudo de caso para 29-30 de outubro de 2003 na região sul do Brasil. In: **III Simpósio Brasileiro de Ciências Geodésicas e Tecnologias da Geoinformação**. Recife, p.1-10, 2010.

MILONE, A. C.; WUENSCHE, C. A.; RODRIGUES, C. V.; JABLONSKI, F. J.; CAPELATO, H. V.; VILAS-BOAS, J. W.; CECATTO, J. R.; VILLELA NETO, T. **Introdução à Astronomia e Astrofísica**. 2003. Disponível em: mtc-m18.sid.inpe.br/col/sid.inpe.br/jeferson/2003/08.14.15.10/doc/curso.pdf. Acesso em: Outubro de 2011.

MOLDWIN, M. **An Introduction to Space Weather**. 1ª. ed.. New York: Cambridge University Press, 2008.

MURALIKRISHNA, A. **Previsão do Índice Geomagnético DST utilizando Redes Neurais Artificiais e Árvores de Decisão**. 2009. Dissertação (Mestrado em Computação Aplicada) - Instituto Nacional de Pesquisas Espaciais.

PASSOS, E.; GOLDSCHMIDT, R. Data Mining: Um Guia Prático. 1ª. ed.. Rio de Janeiro: Elsevier, 2005.

PETRY, A. **Construção do banco de dados para o detector multidirecional de muons - MMD-DB**. 2010. Relatório Técnico - Instituto Nacional de Pesquisas Espaciais.

PETRY, A.; ARAUJO, F. V.; COLPO, M. P.; KATO, C.; BUENO, J.; SILVA, M. R.; VIEIRA, L. R.; KEMMERICH, N.; LAGO, A. D.; SCHUCH, N. J.Data management system for multidirectional muon detector. In: **Twelfth International Congress of the Brazilian Geophysical Society**, Rio de Janeiro, Brasil, 2011.

SAVIAN, J. F.; SILVA, M. R.; DAL LAGO, A.; MUNAKATA, K.; GONZALEZ, W. D.; SCHUCH, N. J. Análise de tempestades geomagnéticas super intensas e de estruturas do meio interplanetário relacionadas, atravpes da observação de raios cósmicos de superfície de alta energia. **Revista Brasileira de Geofísica**, v. 23, n. 2, p. 173-179, 2005.

SHEARER, C. The CRISP-DM Model: the new blueprint for data mining. **Journal of Data Warehousing**, v.5, n.4, p.13-22, 2000.

SILVA, M. R. **Variação da intensidade dos raios cósmicos em resposta a diferentes estruturas magnéticas do meio interplanetário**. 2005. Dissertação (Mestrado em Geofísica Espacial) - Instituto Nacional de Pesquisas Espaciais.

SILVA, M. P. S. Mineração de Dados: Conceitos, Aplicações e Experimentos com WEKA. In: **IV Escola Regional de Informática do Rio de Janeiro e Espírito Santo**. Rio das Ostras

e Vitória, Brasil, 2004.

SOHO - Solar and Heliospheric Observatory. **Gallery - Best of SOHO**. Disponível em: http://sohowww.nascom.nasa.gov/gallery/bestofsoho.html. Acesso em: Outubro de 2011.

University of Waikato. **WEKA 3 - Data Mining Software in Java**. Disponível em: http://www.cs.waikato.ac.nz/ml/weka. Acesso em: Outubro de 2011.

University of Waikato. **Attribute-Relational File Format (ARFF)**. Disponível em: http://www.cs.waikato.ac.nz/ml/weka/arff.html. Acesso em: Outubro de 2011.

WITTEN, I. H.; FRANK, E.; HALL, M. A. **Data Mining: Practical machine learning tools and techniques**. 3ª. ed. San Francisco: Morgan Kaufmann, 2011.

YAMASHITA, C. S. **Efeito das tempestades magnéticas intensas na ionosfera de baixa latitude**. 1999. Dissertação (Mestrado em Geofísica Espacial) - Instituto Nacional de Pesquisas Espaciais.