Busca de Regras para Previsão de Tempestades Geomagnéticas Utilizando Técnica de Mineração de Dados

Amita Muralikrishna
Mestranda em Computação Aplicada
Instituto Nacional de Pesquisas Espaciais
amita.mk@lac.inpe.br

Alisson Dal Lago
Divisão de Geofísica Espacial
Instituto Nacional de Pesquisas Espaciais
dallago@dge.inpe.br

Rafael Duarte Coelho dos Santos
Laboratório Associado de Computação e Matemática Aplicada
Instituto Nacional de Pesquisas Espaciais
rafael.santos@lac.inpe.br

José Demísio Simões da Silva
Laboratório Associado de Computação e Matemática Aplicada
Instituto Nacional de Pesquisas Espaciais
demisio@lac.inpe.br

Resumo

A Terra sofre uma considerável influência da atividade solar através do vento solar, que traz consigo estruturas resultantes, principalmente, de explosões solares ou flares e ejeções de massa coronais - EMC – entre outros eventos que ocorrem na fotosfera (visível) e na coroa solar. O avanço destas estruturas, dependendo de suas características, pode causar tempestades magnéticas na Terra, as quais, por sinal, podem ocasionar diversos danos aos sistemas tecnológicos em nosso planeta.

Uma das soluções para diminuir os estragos causados pelas tempestades magnéticas intensas é a previsão de suas ocorrências, a qual é abordada no presente trabalho.

A técnica escolhida para esta tarefa é a Árvore de Decisão, uma das técnicas de Mineração de Dados.

Palavras-chave: *Tempestade Magnética, Data Mining, Clima Espacial, Geofísica Espacial.*

1. Introdução

Estudos do clima espacial envolvem observações do avanço de estruturas provenientes da coroa solar, que se deslocam pelo meio interplanetário juntamente com o fluxo constante originado no Sol conhecido por vento solar. Quando estas estruturas vêm em direção à Terra, apesar de grande parte ser desviada pela frente de choque formada pelo campo magnético terrestre, formando a magnetosfera, parte de suas energias podem penetrar através dessa barreira do nosso planeta direta ou indiretamente, e conseqüentemente ocasionar mudanças no comportamento do campo magnético e na composição da atmosfera da Terra [2]. Tais alterações podem causar diversos problemas como uma considerável interferência principalmente nos sinais de comunicação e instrumentos de navegação a bordo de satélites e causar alguns fenômenos não visíveis como a geração da corrente de anel e visíveis como as auroras.

Recentes trabalhos estudam a utilização de técnicas de inteligência artificial, como por exemplo, de Redes Neurais Artificiais, para a previsão de índices geomagnéticos [5]. O presente trabalho foca a utilização de técnicas de Mineração de Dados para tentar extrair regras baseadas em dados de plasma e campo magnético interplanetário nas ocorrências de tempestades magnéticas, representadas pelo índice geomagnético Dst.

1.1 Interação Sol-Terra

A Terra sofre constante influência do Sol através do chamado vento solar, constituído por plasma, que se desloca pelo espaço interplanetário com velocidades da ordem de 450 km/s [1]. Eventualmente ocorrem ejeções de plasma solar através de fenômenos explosivos conhecidos como ejeções de massa coronal [4], as quais podem ter velocidades entre 500 e 1500km/s, podendo levar de 2 a 3 dias para cruzar os 150

milhões de quilômetros que separam a Terra do Sol [3] [7].

Grande parte da matéria e energia dessas estruturas é desviada pela magnetosfera terrestre (Figura 1), camada que se forma pela interação entre o vento solar e o campo magnético da Terra.

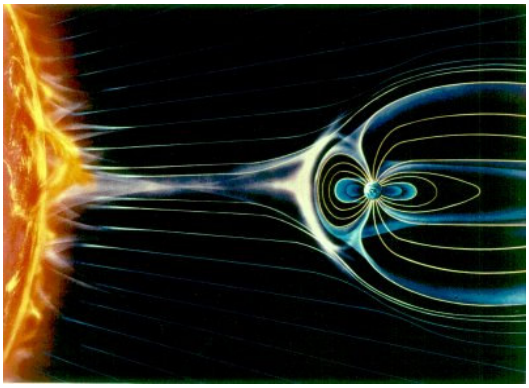

Figura 1: Ilustração do Vento Solar ao se chocar com o campo magnético da Terra.

A energia e a matéria não desviadas pelo campo magnético da Terra podem penetrar no planeta de variadas formas. De acordo com as suas características, estes processos podem ser imperceptíveis para a vida na Terra ou causar fenômenos e mudanças que se tornem visíveis ou até mesmo prejudiciais para as atividades no planeta, o que geralmente ocorre quando o vento traz consigo matéria proveniente de eventos de atividade intensa no Sol. Estes eventos podem ser explosões solares (ou *flares*), ejeções de massa coronais, entre outros, entre os quais a maioria libera matéria ionizada, da qual é constituída a atmosfera solar. A Figura 2 mostra uma erupção solar registrada pelo satélite SOHO - *Solar and Heliospheric Observatory*.

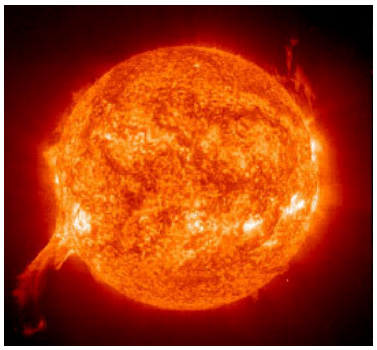

Figura 2: Imagem captada pelo satélite SOHO mostra um flare (canto inferior esquerdo).

Um dos fenômenos causados pela incidência indireta da radiação na Terra é conhecido como aurora (Figura 3), uma iluminação do céu noturno sob a forma de belíssimas "folhas ondulantes" de luz que são vistas nas regiões próximas aos pólos. É resultado da excitação dos átomos e moléculas constituintes dos gases pertencentes à atmosfera superior ao serem atingidos por partículas provenientes do vento solar (elétrons e prótons). Na Terra, a maioria das auroras ocorre a uma altura de 100 a 250 km da superfície.

Figura 3: Aurora sobre o Lago Scwhatka, em Whitehorse, Yukon,nos Estados Unidos.

Além de influir na propagação das ondas de rádio, o vento solar pode influir no comportamento da atmosfera da Terra, pois as partículas carregadas podem alterar a ionização na alta atmosfera, quando há a ocorrência de tempestades magnéticas.

Durante estes períodos ditos "geomagneticamente perturbados", a radiação afeta os equipamentos eletrônicos dos satélites, prejudicando as comunicações. Os próprios satélites podem ser danificados ou perdidos. As camadas superiores da atmosfera se aquecem e se expandem e podem atingir a altura de um satélite. O atrito pode, então, desacelerar o satélite e modificar sua órbita. Em caso de *flares* solares muito intensos, astronautas em órbita correm riscos de morte se forem expostos a chuvas de partículas de alta energia que são aceleradas durantes estes eventos solares. Até passageiros de aviões podem sofrer algum risco.

1.2 Técnicas de mineração de dados ou *data mining*

Nos tempos atuais, um grande volume de dados é gerado, coletado e armazenado. No entanto, pequena parte desses dados é analisada. E assim, importantes informações podem estar sendo perdidas por falta de se ter um método que extraia dos dados todo o conhecimento que eles oferecem.

A mineração de dados é uma área que envolve diversas técnicas que buscam extrair informações relevantes de um grande número de dados. Além de focarem a análise por métodos convencionais,

permitem a descoberta de comportamentos e/ou padrões que não se obteriam se os dados fossem processados independentemente.

A mineração de dados abrange diversas tarefas, entre elas, as de classificação, descoberta de regras e agrupamento.

A técnica utilizada neste trabalho foi a árvore de decisão, mais especificamente, o algoritmo J4.8, disponível no *software Weka* [8]. Trata-se de um procedimento de descoberta de regras, no qual apresenta-se ao algoritmo dados com respostas conhecidas para que através destes sejam geradas regras classificatórias do tipo "se, então", que indicam a classe mais provável para cada registro de dados.

2. Metodologia e desenvolvimento

Tomando-se os fatores prejudiciais descritos nos itens anteriores, a previsão de tempestades magnéticas seria um serviço importante de alerta aos responsáveis da possibilidade da ocorrência do evento.

O objetivo é procurar resultados iniciais para o estudo da previsão de tempestades geomagnéticas, com a análise de dados do meio interplanetário coletados pelo satélite ACE - Advanced Composition Explorer, durante a emissão de matérias carregadas acopladas ao vento solar.

O satélite ACE, situado no ponto lagrangeano (ponto interplanetário entre a Terra e o Sol onde as forças gravitacionais de ambos se compensam), coleta dados de plasma - densidade, velocidade e temperatura - e de campo magnético interplanetário - componente bx, by, bz e a resultante |b|. Os dados de plasma são coletados a cada 64 segundos e os de campo magnético a cada 4 minutos.

Os índices geomagnéticos Dst são medidos na superfície terrestre em várias estações de coleta espalhadas no mundo, sob cuidados de institutos de pesquisas e universidades, sendo coletados de hora em hora.

Após a medição dos dados interplanetários, o evento causador de uma possível supertempestade, pode levar minutos ou até horas, dependendo de como ocorre a sua entrada na atmosfera.

Para este trabalho foram utilizados dois tipos de dados:

- medidas das três componentes do eixo do campo magnético terrestre, que medem as direções que toma a estrutura de plasma ao chocar-se com a magnetosfera da Terra – valores observados com antecedência;
- medidas de índice geomagnético – Dst - que mede as variações que ocorrem no campo magnético terrestre – valores ou classes comportamentais que se deseja prever.

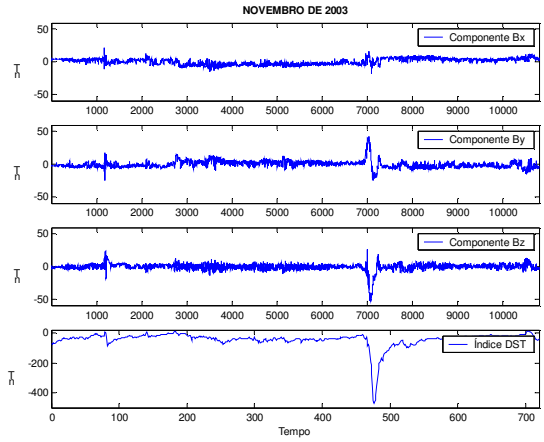

Figura 4: Dados brutos de componentes magnéticas interplanetárias e do índice Dst para o mês completo de Novembro de 2003.

Para este trabalho, foi utilizado um mês completo de dados – Novembro de 2003 – durante o qual foi observada uma super-tempestade. A Figura 4 mostra os dados em brutos utilizados.

Inicialmente tínhamos quinze valores de componentes para um valor de índice, devido aos intervalos de coleta de cada tipo de dados. Foi realizado um pré-processamento para igualar essas escalas para de uma em uma hora.

Nota-se que nos dados brutos já é possível identificar uma modificação do comportamento das componentes em momentos próximos à super-tempestade observada no índice Dst, que ocorre quando o valor do índice cai subitamente chegando a valores inferiores a -400 nT.

Porém, observou-se que a utilização dos dados brutos de campo magnético não seria viável pela presença constante de oscilações não significativas. Para a atenuação dos sinais, foram calculadas a média e a variação sobre cada coluna de dados. Este processamento foi aproveitado também para diminuir o intervalo entre os valores de quatro minutos para uma hora, para que ficassem com o mesmo intervalo dos dados de índice Dst.

Como primeiro passo, foram calculadas médias e variâncias flutuantes, ou seja, cada ponto do vetor teve a média e a variância calculadas em cima dos sete valores anteriores e os sete posteriores. Para se obter um vetor com intervalo de uma hora, foi tirado o valor máximo de variância entre cada conjunto de quinze valores e a média entre cada conjunto de quinze médias. A Figura 5 mostra a atenuação no sinal obtida com o cálculo das médias.

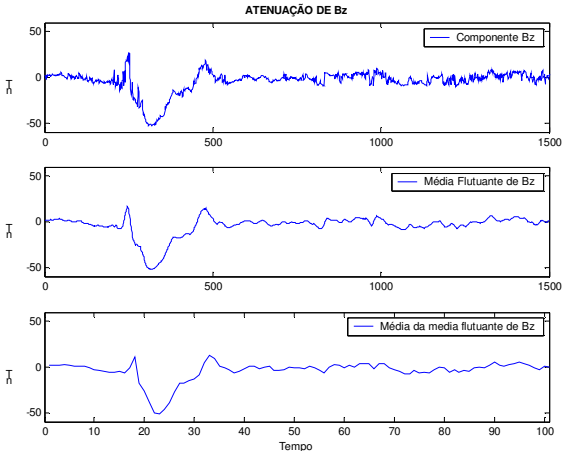

Figura 5: A atenuação do sinal das componentes com o cálculo das médias flutuantes e finais.

Os algoritmos utilizados para o processamento dos dados exigiram que estes estivessem em formato ARFF. A Figura 6 mostra a parte inicial de um dos arquivos ARFF utilizados no na fase de processamento, no qual os atributos são numéricos e separados por vírgula (,) e os rótulos são nominais e vêm no fim de cada registro (linha).

Observa-se na Figura 6 que foram utilizados valores de índice Dst discretizados em classes que serviram de rótulos para os demais atributos. Os valores foram classificados de acordo com a intensidade e utilizados como valores nominais, procedimento o qual é melhor detalhado a seguir.

```
% Arquivo com dados de média e variância da componente bz para a mesma
% e para uma e duas horas anteriores à medição do índice DST.

@relation medias_variancias_bz
@attribute mbz-2 numeric
@attribute mbz-1 numeric
@attribute mbz numeric
@attribute vbz-2 numeric
@attribute vbz-1 numeric
@attribute vbz numeric
@attribute class {fraca,moderada,super}

@data
2.377000,1.701000,0.951000,0.002665,0.040061,0.042172,fraca
1.701000,0.951000,0.700000,0.040061,0.042172,0.162010,fraca
0.951000,0.700000,1.192000,0.042172,0.162010,0.406006,moderada
1.192000,3.123000,2.379000,0.406006,0.400047,0.063086,moderada
1.820000,0.796000,-1.030000,0.078135,0.019728,0.005932,moderada
0.796000,-1.030000,-2.372000,0.019728,0.005932,0.008515,super
-2.372000,0.867000,1.665000,0.008515,0.037257,0.030258,super
-1.748000,-0.563000,-3.357000,0.021495,0.010434,0.016505,moderada
-0.563000,-3.357000,-2.888000,0.010434,0.016505,0.011269,moderada
```

Figura 6: Parte inicial de um arquivo ARFF.

3. Resultados

A tarefa de prever o comportamento do índice geomagnético Dst requer um algoritmo que permita descobrir qual o movimento que ocorre nos eixos da coordenada do campo magnético da Terra no momento em que a magnetosfera se choca com uma estrutura de plasma provinda do Sol. Os dados relativos a essa estrutura também seriam relevantes neste trabalho, porém, por questões de tempo e praticidade, decidiu-se por utilizar somente os dados de campo magnético, já que estes estão todos no mesmo intervalo de tempo, o que não ocorre com os dados de plasma, que se encontram coletados em um outro intervalo de tempo.

Para esse objetivo foi escolhida uma árvore de decisão que permitirá descobrir regras que descrevam o comportamento do índice geomagnético em função dos dados interplanetários.

Optou-se pelo algoritmo J4.8, presente no Weka [8], que trata com atributos numéricos e com classes nominais que rotulam cada linha de dados, ou seja, os dados apresentados ao algoritmo já se encontram classificados. O algoritmo então tenta achar as regras de produção entre os atributos que levaram a aquela separação de classes (ou classificação).

Para os valores dos atributos, foram realizados alguns testes utilizando as três componentes e outros utilizando somente a componente bz, cujo comportamento é considerado um dos principais indicadores da ocorrência de tempestades.

No caso dos rótulos, foram realizados testes com diferentes números de classes.

Os índices Dst, que originalmente são numéricos e coletados de hora em hora, foram discretizados utilizando-se um algoritmo no pré-processamento que cria um vetor do mesmo tamanho que o original com classes, em vez de números. Os valores foram separados em faixas e cada faixa recebeu o nome de uma classe. A Tabela 1 mostra como ficou a conversão de valores numéricos para classes nominais, que consiste no primeiro conjunto de classes:

INTERVALO	CLASSE
$dst > -30$	ausência
$-50 \leq dst \leq -30$	fraca
$-100 \leq dst < -50$	moderada
$-250 \leq dst < -100$	intensa
$dst < -250$	super

Tabela 1 – Primeiro conjunto de classes para os índices Dst

Para apostar em um bom resultado, no primeiro teste foram considerados 19 atributos: para cada componente do eixo, as médias e variâncias nos tempos t, t-1 e t-2 e a classe referente ao tempo t.

Utilizando a configuração J48 –C 0.25 –M 2, foi obtido um resultado satisfatório em termos de porcentagem de erros, porém foi gerada uma árvore muito grande de regras, o que inviabilizou a análise individual de cada regra.

Foi realizado mais um processamento com a mesma configuração para a árvore, porém utilizando como atributos somente os valores de média, não utilizando desta vez, os valores de

variância. A Tabela 2 mostra os resultados que foram muito semelhantes ao primeiro teste, com a diferença, obviamente, do tamanho, já que este apresenta um número menor de atributos.

```
Number of Leaves:  70
Size of the tree :  139
Time taken to build model: 0.22 seconds
=== Evaluation on training set ===
Correctly Classified Instances 601     83.7047 %
Incorrectly Classified Instances 117   16.2953 %
=== Confusion Matrix ===
  a   b   c   d   e  <-- classified as
286  16   6   1   0  |  a = ausencia
 56 187  13   0   1  |  b = fraca
 10   8 105   0   0  |  c = moderada
  2   1   3  13   0  |  d = intensa
  0   0   0   0  10  |  e = super
```

Tabela 2: Avaliação das regras geradas para o teste que utiliza somente valores de média.

Nos dois testes citados acima percebeu-se que as regras são iniciadas sempre com o valor de bz como base. Isso comprova a teoria de que essa componente é um dos parâmetros que mais colaboram na previsão do índice Dst. Mais um fator interessante observado nos resultados é o erro mínimo na verificação das regras que classificam as super-tempestades, cuja previsão possui um grau de importância superior aos demais eventos. O início da árvore de regras de um dos dois primeiros testes citados acima, representada na Tabela 3 mostra a classificação correta de 10 de 11 ocorrências de super-tempestades e a tomada do atributo *bz* como base das regras.

```
J48 pruned tree
------------------
mbz-2 <= -6.04952
|  mbx-2 <= 1.844782: super (11.0/1.0)
|  mbx-2 > 1.844782
|  |  mbx-2 <= 5.354289: moderada (7.0/1.0)
|  |  mbx-2 > 5.354289: intensa (3.0)
mbz-2 > -6.04952
|  mbx-1 <= -1.983987
|  |  mbz-1 <= 3.348733
|  |  |  mbz-2 <= 0.933556
|  |  |  |  mbx <= -3.822289
```

Tabela 3: Início da árvore de regras de um dos primeiros testes.

Pode-se considerar essas regras um resultado já satisfatório. Porém, o que impede que seja dada uma grande significância a estes resultados é o fato de que o teste foi realizado somente para um mês de dados, no qual só ocorre uma super-tempestade. O ideal será a realização de testes com mais meses com ocorrências de tempestades magnéticas, para a verificação da validade destas regras para os demais casos. São testes que ficarão como planos futuros para a continuidade deste trabalho.

Outro teste realizado para o mesmo conjunto de regras foi a utilização somente dos valores de bz como atributos. Notando que a componente bz influencia de forma significativa na ocorrência ou não de tempestades magnéticas, foram realizados testes nos quais foram utilizados como atributos os valores relativos à componente bz. Porém, estes testes apresentaram resultados ruins em relação aos primeiros e, portanto não serão descritos mais detalhadamente.

Analisando as matrizes de confusão geradas nos testes anteriores, a aposta seguinte foi tentar utilizar um segundo conjunto de classes com menos classes que o primeiro conjunto. Observou-se a mistura entre as classes 'ausência e fraca' e entre 'fraca e moderada'. A idéia é tentar achar uma forma de unir essas classes sem prejudicar os critérios das classes mais significantes, que são a 'intensa' e a 'super'. Seguem na Tabela 4 os novos critérios de classificação.

INTERVALO	CLASSE
$dst > -70$	fraca
$-150 < dst \leq -70$	moderada
$dst < -150$	super

Tabela 4 - Segundo conjunto de classes para os índices Dst

Realizando os mesmos testes que foram realizados para o primeiro conjunto de regras, temos primeiro a árvore de decisão gerada para os 19 atributos entre eles as médias e variâncias das três componentes em três momentos no tempo. A Tabela 5 mostra as suas características.

```
Number of Leaves:  17
Size of the tree:  33
=== Evaluation on training set ===
Correctly Classified Instances 698     97.2145 %
Incorrectly Classified Instances 20     2.7855 %
=== Confusion Matrix ===
  a   b   c  <-- classified as
641   4   1  |  a = fraca
  8  43   1  |  b = moderada
  5   1  14  |  c = super
```

Tabela 5: Avaliação das regras geradas pelo algoritmo para o terceiro teste

Nota-se que este foi o melhor teste em termos de porcentagem de acertos de classificação, com 97%. Porém, se formos avaliar o número de

classificações errôneas para a classe de super-tempestades, os primeiros testes são melhores.

O teste com a utilização somente de valores de médias como atributos apresentou um resultado muito semelhante a este citado acima, como ocorreu também com os dois primeiros testes com o primeiro conjunto de classes. E, portanto, não entraremos em detalhes em relação a ele.

O teste com as médias e variâncias de bz, como ocorrido para o primeiro conjunto de classes, não apresentou resultados bons.

3.1 Validação

A fim de analisar como um conjunto de regras já criado se comporta com dados novos de teste, salvamos o conjunto de regras do primeiro teste realizado com o primeiro conjunto de classes, considerado o melhor teste por ter classificado corretamente quase todas as super-tempestades (90.9% de acerto) e tomamos um novo conjunto de dados, relativo ao período do mês de Outubro de 2003 (Figura 7), e testamos com as regras já geradas.

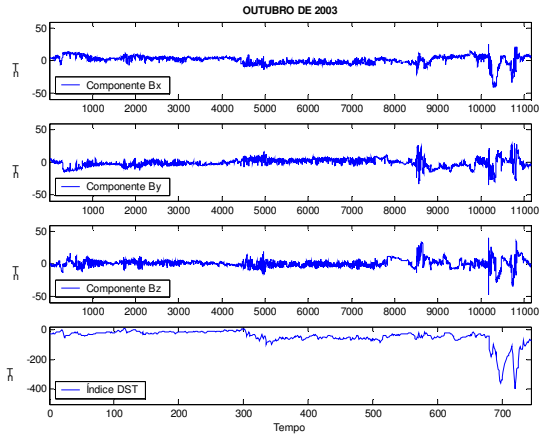

Figura 7: Dados brutos de Outubro de 2003.

A Tabela 6 mostra a qualidade da classificação obtida na validação.

=== Confusion Matrix ===					
a	b	c	d	e	<-- classified as
200	72	46	13	1	| a = ausencia
69	54	34	9	1	| b = fraca
80	60	34	6	9	| c = moderada
18	8	9	1	0	| d = intensa
3	2	1	0	12	| e = super

Tabela 6: Matriz de Confusão gerada na Validação.

A validação retornou um resultado não-confiável já que boa parte das super-tempestades foram classificadas erroneamente. Para a obtenção de melhores resultados, sugere-se para trabalhos futuros a utilização de mais meses de dados com ocorrências de tempestades para a fase de geração de regras.

4. Conclusão

A árvore de decisão mostrou-se um algoritmo muito promissor tratando-se da aplicação na previsão de tempestades magnéticas, pois mesmo em uma árvore com muitos nós, poucas regras podem oferecem grande confiança nos resultados e conseqüentemente qualidade na classificação.

É importante salientar que em nenhum dos testes foi realizada uma análise completa de todas as regras, pois se trata de um número grande de regras e não seria o objetivo do projeto a realização de uma análise minuciosa. No entanto, na fase de continuidade deste trabalho pretende-se fazer uma revisão de todo o comportamento da árvore de decisão.

Chegou-se à conclusão, através do teste de validação, que o comportamento de um só mês de dados não é o suficiente para a descoberta de regras confiáveis se tratando da previsão geral de super-tempestades, isto é, para qualquer ocorrência deste fenômeno. A utilização de mais meses de dados para a geração de regras talvez trará resultados mais confiáveis na validação com qualquer evento do tipo.

4.1 Planos futuros e melhorias

Para a mesma aplicação deseja-se futuramente realizar mais testes utilizando árvore de decisão (algoritmo J 4.8) com diferentes conjuntos de atributos e com maior volume de dados observados. E juntamente, pretende-se testar outros algoritmos de classificação de *data mining*.

Outra sugestão para trabalhos futuros é a previsão de uma tempestade completa, isto é, toda a sua duração, em vez de se prever somente um valor de índice (representado por valor numérico ou classe) isoladamente, o que seria interessante pra prever o comportamento de índices em duas ou mais horas à frente em vez de uma única hora.

5. Referências

[1] Brant, J. C. *Introduction to the solar wind*. San Francisco: W. H. Freeman, 1970. 216p.

[2] Cowley, W. H. *A Beginner's gide to the Earth's Magnetosphere*. EOS, Dec. 19, p.525, 1995.

[3] Gosling, J. T.. *Coronal mass ejections: the link between solar and geomagnetic activity*. Phys. Fluids B 5, v.7, p.2638-2645, July 1993

[4] Hunduhausen, A. J. *An introduction*. In: Crooker, N.; Joselyn J. A.; Feynman, J. ed. Coronal mass ejections, Washington, DC: AGU, 1997. v. 99, p.1-7.

[5] Lundstedt, H.. *AI Techniques in Geomagnetic Storm Forecasting*, invited review paper in Magnetic Storms, Geophysical Monograph 98, AGU, 1997.

[6] Rafael Santos, *Material didático* - - http://www.lac.inpe.br/~rafael.santos/

[7] Schwenn, Rainer; Dal Lago, A.; Huttenen, Emilia & Gonzalez, Walter D.. *The association of coronal mass ejection with their effects near the Earth*. Annales Geophysicae-Atmospheres Hydrospheres And Space Sciences, v. 23, n. AG/2004180, p. 1033-1059, 2005.

[8] Software Weka - http://www.cs.waikato.ac.nz/ml/weka