AVALIAÇÃO DE ÁRVORES DE DECISÃO PARA DATA MINING DE DADOS METEOROLÓGICOS

José Carlos Becceneri
Laboratório Associado de Computação e Matemática Aplicada – INPE
São José dos Campos, SP, Brasil, 12227-010
becce@lac.inpe.br

Yasuo Kono
Divisão de Sistemas Solo – INPE
São José dos Campos, SP, Brasil, 12227-010
yasuo@dss.inpe.br

Rubens Cruz Gatto
Divisão de Sistemas Solo – INPE
São José dos Campos, SP, Brasil, 12227-010
gatto@dss.inpe.br

Rafael Santos
Instituto de Estudos Avançados – Centro Técnico Aeroespacial
São José dos Campos, SP, Brasil, 12228-840
rafael@ieav.cta.br

Resumo

Técnicas de *Data Mining* podem ser usadas para analisar dados em uma base de dados para descrever as relações existentes entre os dados ou para prever o comportamento de dados que possam ser ou não parte daquela base. O objetivo de nossa pesquisa atual é tentar descrever o comportamento dos dados contidos em uma base de dados meteorológicos para que através da descrição dos dados como regras possamos identificar exceções ou *outliers*.

Neste artigo apresentamos alguns resultados experimentais da extração e classificação de regras obtidas de bases de dados meteorológicos usando determinadas métricas e técnicas de *Data Mining*. Observamos experimentalmente que a complexidade dos dados impede uma descrição com as técnicas apresentadas.

Palavras-Chaves: Data Mining, dados meteorológicos, árvores de decisão.

Abstract

Data mining techniques can be used to analyze databases in order to describe existing relationships between the data or to predict the behavior of data which can belong or not to the database. The purpose of our research is to attempt to describe the behavior of the data in a meteorological database so through the description of the data as production rules outliers can be found.

In this article we present some experimental results on the extraction and classification of the rules obtained from the meteorological databases using certain data mining metrics and techniques. We experimentally observe that the complexity of the data makes the data description with the presented techniques a very complex task.

Keywords: Data Mining, meteorological database, decision trees.

1. INTRODUÇÃO

Uma grande quantidade de dados meteorológicos tem sido coletada por centenas de estações meteorológicas de solo chamadas Plataformas de Coleta de Dados (PCDs) distribuídas por

todo o Brasil. Estes dados são transmitidos aos satélites de coleta de dados SCD1 e SCD2 e enviados para estações de solo em Cuiabá e Alcântara, que os encaminham para o CPTEC – Centro de Previsão de Tempo e Estudos Climáticos. Os dados são pré-processados (calibrados) e disponibilizados para usuários finais.

No sistema de coleta de dados, os dados ambientais são avaliados quanto à qualidade e confiabilidade através da análise do protocolo de transmissão das mensagens transmitidas pelas PCDs e recebidas no centro de processamento. Esse processo garante a conformidade do dado recebido, e disponibilizado ao usuário final, com o dado que foi medido pelo sensor "in loco". Porém, não é atribuição do sistema de coleta de dados fazer correlações espaciais e temporais entre os valores dos parâmetros ambientais medidos. Essa é uma atividade que apenas o usuário final, especialista, pode realizar. As técnicas apresentadas neste artigo visam fornecer ao usuário final especialista ferramentas para a análise e correlação dos dados recebidos para sua validação final.

Neste artigo avaliaremos algumas métricas para classificação de regras que podem ser obtidas a partir de bancos de dados com técnicas de *Data Mining*. O nosso objetivo é classificar as regras em relação a estas métricas para verificar se as mesmas descrevem adequadamente os dados na base de dados, e, a partir disso, tentar futuramente identificar inconsistências nos dados. É importante ressaltar que o objetivo desta pesquisa **não** é descobrir ou descrever regras que permitam a previsão do tempo com qualquer precisão, mas tão somente descobrir regras e exceções que indiquem possíveis problemas no processo de coleta dos dados.

Este artigo está dividido nas seguintes seções: a seção 2 apresenta os conceitos de *Data Mining* pertinentes à análise proposta. A seção 3 apresenta informações sobre as plataformas de coleta de dados e seus sensores. A seção 4 apresenta um experimento de *Data Mining* feito com dados de uma estação PCD e a seção 5 apresenta conclusões e direções para pesquisa futura.

2. CLASSIFICAÇÃO DE REGRAS COM TÉCNICAS DE *DATA MINING*

Uma definição concisa que podemos adotar de *Data Mining* é *"a pesquisa por informações valiosas em grandes volumes de dados"* [1]. Essa definição aplica-se bem ao contexto deste trabalho. Possuímos uma grande massa de dados (alguns gigabytes, com crescimento mensal de cerca de 40 megabytes) com muitas informações potencialmente valiosas não exploradas até o início desta nossa pesquisa. Os dados foram obtidos de plataformas de coleta de dados conforme descrito na introdução deste artigo e detalhado na seção 3 do mesmo.

Dois dos objetivos que podem ser atingidos com *Data Mining* são predizer comportamentos futuros para tomada de decisões e descobrir padrões previamente desconhecidos de comportamentos. Na presente pesquisa estamos interessados no segundo objetivo: usaremos técnicas de *Data Mining* para extrair informações desta massa de dados para tentar localizar exceções a regras aplicáveis aos dados. Estas exceções podem possivelmente caracterizar incoerências nos dados e na forma de coleta dos mesmos.

Diversas técnicas computacionais podem ser empregadas na pesquisa por informações valiosas. Segundo Thearling [2], as técnicas mais comumente usadas de *Data Mining* são:

- Redes Neurais Artificiais [3]
- Árvores de Decisões [4]
- Algoritmos Genéticos [5]

- Método do Vizinho Mais Próximo [6]
- Regras de Indução [7]

Segundo Parpinelli [8], uma técnica que ainda é um campo de pesquisa inexplorado é a técnica Ant Colony [9]. Uma possível continuidade deste trabalho empregará tal técnica para uma análise mais detalhada dos dados em questão.

Para esta pesquisa usaremos a criação de árvores de decisão, a extração de regras de classificação a partir das árvores de decisão e a análise de algumas métricas aplicáveis a estas regras. Os procedimentos em detalhes são apresentados na seção 4 deste artigo.

Explicaremos a seguir, as métricas que iremos utilizar em nosso trabalho conforme definidas por de la Iglesia [10]. Para tanto, vamos considerar uma base de dados **D** e uma regra de classificação que possa ser aplicada a seus registros para classificação. Uma regra de classificação tem a forma *se X então Y*, onde *X* representa um conjunto de testes sobre os atributos de **D** e *Y* é chamado de conseqüente da regra. Freqüentemente *Y* é um outro atributo existente de **D** ou pode ser derivado de outros atributos.

Chamemos de **A** o conjunto de registros de dados para os quais o antecedente *X* pode ser aplicado (isto é, o conjunto de registros para os quais o antecedente *X* é avaliado como verdadeiro, independentemente do conseqüente), **B** o conjunto de registros para os quais o conseqüente é aplicável (ou o conjunto de registros para os quais o conseqüente é *Y*, independentemente do antecedente) e **C** o conjunto de registros para os quais tanto o antecedente quando o conseqüente das regras são aplicáveis e verdadeiros. Evidentemente, $|\mathbf{C}| \leq |\mathbf{A}|$ e $|\mathbf{C}| \leq |\mathbf{B}|$.

A partir destes conjuntos **A**, **B**, **C** e **D**, e considerando a cardinalidade dos mesmos (**a**, **b**, **c** e **d**) podemos descrever as seguintes métricas para uma regra *r* (todas as métricas são medidas em porcentagens):

- **Acurácia** ou **Confiança**, denotada por *Acc(r)* = **c/a**: esta medida corresponde ao percentual dos registros para os quais a predição da regra está correta, tomado sobre o total de registros para os quais o antecedente é aplicável e correto.
- **Aplicabilidade**, denotada por *App(r)* = **a/d**: esta medida representa o percentual de registros no banco de dados que podem ser avaliados por esta regra, ou seja, o percentual de registros para os quais o antecedente da regra é avaliado como sendo verdadeiro.
- **Suporte**, denotado por *Sup(r)* = **c/d**: esta medida, numericamente igual à aplicabilidade multiplicada pela acurácia, corresponde ao percentual de registros que são classificados corretamente quanto ao antecedente e ao conseqüente, em relação a todos os registros do banco de dados. Esta medida é freqüentemente usada com a medida de acurácia ou confiança para estabelecer a qualidade das regras individuais.
- **Cobertura**, denotada por *Cov(r)* = **c/b**: esta medida corresponde ao percentual dos registros que é classificado corretamente pela regra.
- **Acurácia Padrão**, denotada por *DefAcc(classe)* = **b/d**: definida como a proporção dos registros do banco de dados que podem ser classificados com aquele conseqüente. Esta medida é tomada classe a classe e é igual para todas as regras que tenham o mesmo conseqüente.

Um exemplo simples pode ilustrar o cálculo e aplicabilidade destas métricas. Consideremos

um pequeno banco de dados que contém somente uma tabela, mostrada na Tabela 1. A tabela mostra registros de médias de notas obtidas no vestibular e nos primeiros e segundos semestres de um curso superior, com cada registro correspondendo às notas de um determinado aluno. A tabela também mostra se o aluno recebe bolsa ou não da instituição.

Tabela 1 – Exemplo de dados para cálculo das métricas (dados simulados).

Nota Vestibular	Média Notas 1	Média Notas 2	Bolsa
94	84	88	**S**
74	87	81	**N**
48	89	41	**N**
86	81	81	**S**
49	88	61	**N**
81	81	82	**S**
43	55	41	**S**
92	96	98	**S**
87	47	42	**N**
59	54	67	**N**

Consideremos que uma análise dos registros da tabela foi feita e três regras foram obtidas:

- Se a nota do vestibular (*NV*) for maior que 90 o aluno receberá bolsa.
- Se a média de notas do primeiro semestre (*MN1*) for menor que 80 o aluno não receberá bolsa.
- Se a nota do vestibular (*NV*) e as médias dos dois primeiros semestres (*MN1* e *MN2*, respectivamente) forem maior que 80 o aluno receberá bolsa.

As regras, seus antecedentes, e conseqüentes, os tamanhos dos conjuntos **A**, **B**, **C** e **D** e as métricas calculadas são mostradas na Tabela 2. Nesta tabela, *X* indica o antecedente e *Y* o conseqüente da regra. Se o antecedente é composto de várias condições, as mesmas são combinadas com um "e" lógico, representado pelo símbolo **&**.

Tabela 2 – Métricas calculadas para os dados simulados.

Regra	*X*	*Y*	A	B	C	D	*Acc*	*App*	*Sup*	*Cov*	DefAcc
1	NV > 90	S	2	5	2	10	100.0	20.0	20.0	40.0	50.0
2	MN1 < 80	N	3	5	2	10	66.7	30.0	20.0	40.0	50.0
3	NV > 80 & MN1 > 80 & MN2 > 80	S	4	5	4	10	100.0	40.0	40.0	80.0	50.0

Podemos observar nos resultados mostrados na Tabela 2 que a regra 1 tem acurácia de 100%, ou seja, todos registros que correspondem ao antecedente *NV>90* são classificados corretamente como sendo de alunos bolsistas. A regra 2 mostra acurácia de 67%, pois somente 2 dos 3 casos onde a nota *MN1* é menor que 80 são classificados corretamente.

A regra 3 mostra cobertura de 80%, pois dos cinco casos nos quais o conseqüente é *S*, quatro são explicados (previstos corretamente) por esta regra. O suporte para a mesma regra é 40%, pois quatro dos dez registros apresentam conseqüente e antecedente previstos corretamente pela regra. Esta pode ser considerada uma boa regra, que descreve bem os registros que podem ser classificados com ela, mesmo que seja aplicável a poucos dos registros do banco

de dados.

3. DADOS USADOS NO EXPERIMENTO

Os dados sendo analisados são aqueles oriundos das plataformas de coleta de dados (PCDs), transmitidos aos satélites, retransmitidos às antenas receptoras e finalmente pré-processadas em uma rotina chamada calibração para armazenamento e disponibilização aos usuários.

Uma plataforma de coleta de dados é uma estrutura composta de vários sensores que registram informações meteorológicas (temperatura, pressão, direção e velocidade do vento, umidade, etc.). É preparada para operar autonomamente e enviar essas informações utilizando os satélites como via de transmissão (ponte) aos centros de processamento. Equipada de antena, transmissor, bateria, painel solar e principalmente sensores, é utilizada em locais remotos de difícil acesso.

Alguns dos dados coletados pelas PCDs, seus atributos e outras informações são mostrados na Tabela 3.

Tabela 3 – Descrição de alguns dados coletados pelas PCDs.

Atributo	Sigla	Unidade	Freqüência de Coleta
Temperatura do Ar	TempAr	°C	Valor instantâneo a cada 3 horas.
Umidade Relativa do Ar	UmidRel	%	Valor instantâneo a cada 3 horas.
Pressão Barométrica	PressaoAtm	mB	Valor instantâneo a cada 3 horas.
Velocidade do Vento	VelVento	m/s	Valor a cada 3 horas, calculado da média de 200 amostras com 3 segundos de intervalo, 10 minutos antes de cada 3 horas.
Velocidade Máxima do Vento (Rajada)	VelVentoMax	m/s	Valor máximo (rajada) a cada 3 horas, amostras a cada 3 segundos.
Radiação Solar Global	RadSolAcum	MJ/m^2	Valor acumulado a cada 3 horas, integração de 1080 amostras de 10 segundos de intervalo.
Precipitação Acumulada	Pluvio	mm	Valor acumulado mensal a cada 3 horas (o acumulador é zerado automaticamente todo dia 01 de cada mês).

4. EXPERIMENTOS E RESULTADOS

O experimento para avaliar as métricas com os dados coletados pelas PCDs consiste na criação de uma árvore de decisão com dados pré-processados, na criação de regras a partir da árvore de decisão e no cálculo e avaliação das métricas.

Os dados coletados das plataformas de coleta de dados foram pré-processados para criação de uma série temporal ininterrupta (isto é, uma série temporal na qual não faltam valores para os atributos).

Os atributos usados no principal experimento com série temporal são a pressão atmosférica (*PressaoAtm*), a radiação solar global (*RadSolAcum*), a temperatura do ar (*TempAr*), a umidade relativa (*UmidRel*) e a velocidade máxima do vento (*VelVentoMax*). Todos estes atributos foram medidos considerando-se a medida atual (não no sentido de presente, mas

correspondente a um registro qualquer na base de dados), uma tomada três horas antes e uma tomada seis horas antes, em um total de 15 atributos para o antecedente das regras. Um indicador de tempo da medida foi anexado aos nomes dos atributos para melhor identificá-los (por exemplo, *PressaoAtm0*, *PressaoAtm3*, *PressaoAtm6*, etc.).

O pré-processamento também criou artificialmente um registro que será considerado o conseqüente para as regras extraídas. Este conseqüente é calculado como a existência ou não de precipitação (diferença entre a precipitação acumulada) entre seis e nove horas após o registro sendo considerado. Este atributo será representado por **S** e **N**, correspondentes, respectivamente, à ocorrência ou não de precipitação. O procedimento para cálculo de precipitação é similar ao feito por um pluviógrafo [11].

A árvore de decisão foi criada usando os dados pré-processados e algoritmo J4.8 do software Weka [12]. Vários parâmetros podem ser usados para configurar a criação de árvores de decisão, para este experimento considerou-se a criação de uma árvore sem ser podada, isto é, sem ter folhas correspondentes a regras agrupadas para simplificar o conjunto de regras às custas de causar erros de classificação.

A árvore de decisão criada com os dados do experimento principal é mostrada na Figura 1.

```
PressaoAtmMinus6 <= 952
 | PressaoAtmMinus0 <= 954
 | | UmidRelMinus0 <= 92
 | | | RadSolAcumMinus3 <= 3.6: no (43/1)
 | | | RadSolAcumMinus3 > 3.6
 | | | | RadSolAcumMinus6 <= 7.6
 | | | | | PressaoAtmMinus3 <= 946: yes (2)
 | | | | | PressaoAtmMinus3 > 946
 | | | | | | RadSolAcumMinus3 <= 4.8
 | | | | | | | RadSolAcumMinus3 <= 4.1
 | | | | | | | | TempArMinus6 <= 17.5: yes (2)
 | | | | | | | | TempArMinus6 > 17.5: no (2)
 | | | | | | | RadSolAcumMinus3 > 4.1: yes (2)
 | | | | | | RadSolAcumMinus3 > 4.8: no (5)
 | | | | RadSolAcumMinus6 > 7.6: no (6)
 | | UmidRelMinus0 > 92
 | | | RadSolAcumMinus3 <= 2.3
 | | | | PressaoAtmMinus0 <= 950: no (23/2)
 | | | | PressaoAtmMinus0 > 950
 | | | | | UmidRelMinus6 <= 78: no (13/2)
 | | | | | UmidRelMinus6 > 78
 | | | | | | VelVentoMaxMinus0 <= 4.3: yes (6)
 | | | | | | VelVentoMaxMinus0 > 4.3
 | | | | | | | VelVentoMaxMinus3 <= 5.4: no (5/1)
 | | | | | | | VelVentoMaxMinus3 > 5.4: yes (5)
 | | | RadSolAcumMinus3 > 2.3: yes (3)
 | PressaoAtmMinus0 > 954: no (11)
PressaoAtmMinus6 > 952: no (163/7)
```

Figura 1 – Árvore de decisão obtida com os dados do experimento principal.

A árvore de decisão mostrada na Figura 1 é mostrada como texto. A sua interpretação é simples, os antecedentes são mostrados cumulativamente em diferentes níveis de indentamento (com conjuntos de caracteres "|" indicando o nível do indentamento), e todas as linhas com conseqüentes mostram um ou dois números entre parênteses. Se houver um número, esse corresponde ao número de registros que foi classificado com aquela regra. Se houver um segundo número, separado do primeiro por uma barra, este indicará o número de

registros que foi classificado incorretamente com aquela regra.

Como um exemplo, as quatro primeiras linhas mostradas na Figura 1 podem ser interpretadas como "Se a pressão atmosférica há seis horas atrás foi menor ou igual a 952 milibares e a pressão atmosférica atual for menor que 954 milibares e a umidade relativa atual for menor ou igual a 92% e a radiação solar acumulada há três horas atrás for menor que 3.6 MJ/m^2 então não houve precipitação entre seis e nove horas após o momento atual". Esta regra classificou 43 registros corretamente e um registro incorretamente.

Outras regras podem ser inferidas da mesma forma. A árvore de decisão possibilita a derivação de 15 regras para os dados deste experimento.

O algoritmo cria, além da árvore de decisão, uma matriz de confusão que indica o número de registros que foi classificado correta e incorretamente em cada classe prevista. A matriz de confusão para esta árvore de decisão indica que dos 291 registros considerados, 20 foram classificados corretamente como ocorrência de precipitação, 258 foram classificados corretamente como falta de precipitação e 13 foram previstos como precipitação, mas classificados incorretamente como falta de precipitação.

A árvore de decisão mostrada na Figura 1 foi usada para criação das regras e cálculo das métricas estudadas neste artigo. As regras e métricas são mostradas na Tabela 4. Para melhor visualização das regras, as mesmas foram ordenadas de forma decrescente pelo valor da métrica Suporte e a métrica Acurácia Padrão foi omitida (o valores desta métrica são 88.7% para as regras cujo conseqüente é **N** e 11.3% para as regras cujo conseqüente é **S**).

Tabela 4 – Regras e métricas obtidas da árvore de decisão.

Regra	***X***	***Y***	**A**	**B**	**C**	**D**	***Acc***	***App***	***Sup***	***Cov***
1	PressaoAtm6 > 952	N	163	258	156	291	95.7	56.0	53.6	60.5
2	PressaoAtm6 ≤ 952 & PressaoAtm0 ≤ 954 & UmidRel0 ≤ 92 & RadSolAcum3 ≤ 3.6	N	43	258	42	291	97.7	14.8	14.4	16.3
3	PressaoAtm6 ≤ 952 & PressaoAtm0 ≤ 954 & UmidRel0 > 92 & RadSolAcum3 ≤ 2.3 & PressaoAtm0 ≤ 950	N	23	258	21	291	91.3	7.9	7.2	8.1
4	PressaoAtm6 ≤ 952 & PressaoAtm0 ≤ 954 & UmidRel0 > 92 & RadSolAcum3 ≤ 2.3 & PressaoAtm0 > 950 & UmidRel6 ≤ 78	N	13	258	11	291	84.6	4.5	3.8	4.3
5	PressaoAtm6 ≤ 952 & PressaoAtm0 > 954	N	11	258	11	291	100.0	3.8	3.8	4.3
6	PressaoAtm6 ≤ 952 & PressaoAtm0 ≤ 954 & UmidRel0 ≤ 92 & RadSolAcum3 > 3.6 & RadSolAcum6 > 7.6	N	6	258	6	291	100.0	2.1	2.1	2.3
7	PressaoAtm6 ≤ 952 &	S	6	33	6	291	100.0	2.1	2.1	18.2

	PressaoAtm0 ≤ 954 & UmidRel0 > 92 & RadSolAcum3 ≤ 2.3 & PressaoAtm0 > 950 & UmidRel6 > 78 & VelVentoMax0 ≤ 4.3									
8	PressaoAtm6 ≤ 952 & PressaoAtm0 ≤ 954 & UmidRel0 ≤ 92 & RadSolAcum3 > 3.6 & RadSolAcum6 ≤ 7.6 & PressaoAtm3 > 946 & RadSolAcum3 > 4.8	N	5	258	5	291	100.0	1.7	1.7	1.9
9	PressaoAtm6 ≤ 952 & PressaoAtm0 ≤ 954 & UmidRel0 > 92 & RadSolAcum3 ≤ 2.3 & PressaoAtm0 > 950 & UmidRel6 > 78 & VelVentoMax0 > 4.3 & VelVentoMax3 ≤ 5.4	N	5	258	4	291	80.0	1.7	1.4	1.6
10	PressaoAtm6 ≤ 952 & PressaoAtm0 ≤ 954 & UmidRel0 > 92 & RadSolAcum3 ≤ 2.3 & PressaoAtm0 > 950 & UmidRel6 > 78 & VelVentoMax0 > 4.3 & VelVentoMax3 > 5.4	S	5	33	5	291	100.0	1.7	1.7	15.2
11	PressaoAtm6 ≤ 952 & PressaoAtm0 ≤ 954 & UmidRel0 > 92 & RadSolAcum3 > 2.3	S	3	33	3	291	100.0	1.0	1.0	9.1
12	PressaoAtm6 ≤ 952 & PressaoAtm0 ≤ 954 & UmidRel0 ≤ 92 & RadSolAcum3 > 3.6 & RadSolAcum6 ≤ 7.6 & PressaoAtm3 ≤ 946	S	2	33	2	291	100.0	0.7	0.7	6.1
13	PressaoAtm6 ≤ 952 & PressaoAtm0 ≤ 954 & UmidRel0 ≤ 92 & RadSolAcum3 > 3.6 & RadSolAcum6 ≤ 7.6 & PressaoAtm3 > 946 & RadSolAcum3 ≤ 4.8 & RadSolAcum3 ≤ 4.1 & TempAr6 ≤ 17.5	S	2	33	2	291	100.0	0.7	0.7	6.1
14	PressaoAtm6 ≤ 952 & PressaoAtm0 ≤ 954 & UmidRel0 ≤ 92 & RadSolAcum3 > 3.6 & RadSolAcum6 ≤ 7.6 &	N	2	258	2	291	100.0	0.7	0.7	0.8

	PressaoAtm3 > 946 & RadSolAcum3 ≤ 4.8 & RadSolAcum3 ≤ 4.1 & TempAr6 > 17.5									
15	PressaoAtm6 ≤ 952 & PressaoAtm0 ≤ 954 & UmidRel0 ≤ 92 & RadSolAcum3 > 3.6 & RadSolAcum6 ≤ 7.6 & PressaoAtm3 > 946 & RadSolAcum3 ≤ 4.8 & RadSolAcum3 > 4.1	S	2	33	2	291	100.0	0.7	0.7	6.1

As regras listadas na Tabela 4 mostram além das regras em si (antecedente e conseqüente) obtidas com a árvore de decisão, o número de registros de cada um dos conjuntos (**A**, **B**, **C** e **D**) e os valores das métricas calculadas. Como as linhas da tabela estão ordenadas pela métrica Suporte, as regras apresentadas primeiro são as consideradas mais significativas para este estudo.

Algumas conclusões interessantes que podem ser extraídas dos dados na Tabela 4 são:

- No experimento mostrado e em outros realizados (variando dados, parâmetros para o algoritmo de criação das árvores e abrangência temporal dos dados criados após o pré-processamento) notou-se que regras mais simples como, por exemplo, a regra número 1, podem ser aplicadas a mais de 50% dos registros na base de dados e nestes casos com acurácia superior a 90%. Em todos os experimentos realizados verificou-se que regras simples com grande acurácia correspondem sempre ao conseqüente **N** (não houve precipitação em período posterior).
- Os dados são complexos e árvores de decisão, mesmo quando podadas, criam muitos galhos para conter subdivisões dos valores de atributos (vide regras 14 e 15 na Tabela 4).
- Regras cujo conseqüente é **S** (houve precipitação em período posterior) são geralmente complexas – a medida de complexidade de uma regra é relacionada com o número de componentes no seu antecedente. Uma explicação é que a quantidade de registros cujo conseqüente foi calculado como **S** é pequeno em relação ao número de registros com conseqüente **N**, problema exacerbado pela complexidade das regras e impossibilidade de separar as classes **S** e **N** linearmente.

A possibilidade de separação linear dos dados em conjuntos que contenham cada um somente dados de uma classe é central ao problema sendo estudado, uma vez que uma árvore de decisão gera regras que nada mais são do que pequenos classificadores lineares agrupados (os antecedentes da regra). Se os dados não são separáveis linearmente com nenhuma combinação dos atributos (coletados dos sensores e avaliados ao longo do tempo) pode-se esperar que uma árvore completa, ou seja, uma árvore que classifica todos os dados corretamente, será necessariamente complexa. Em outras palavras, para classificar dados que não são separáveis linearmente (ou que precisem de um grande número de atributos para ser) deveremos criar árvores com um grande número de nós e galhos.

A Figura 2 mostra como os dados não podem ser separados linearmente nem por regras com grande acurácia. Na Figura 2 temos os dados do atributo *PressaoAtm6* plotados contra o conseqüente das regras.

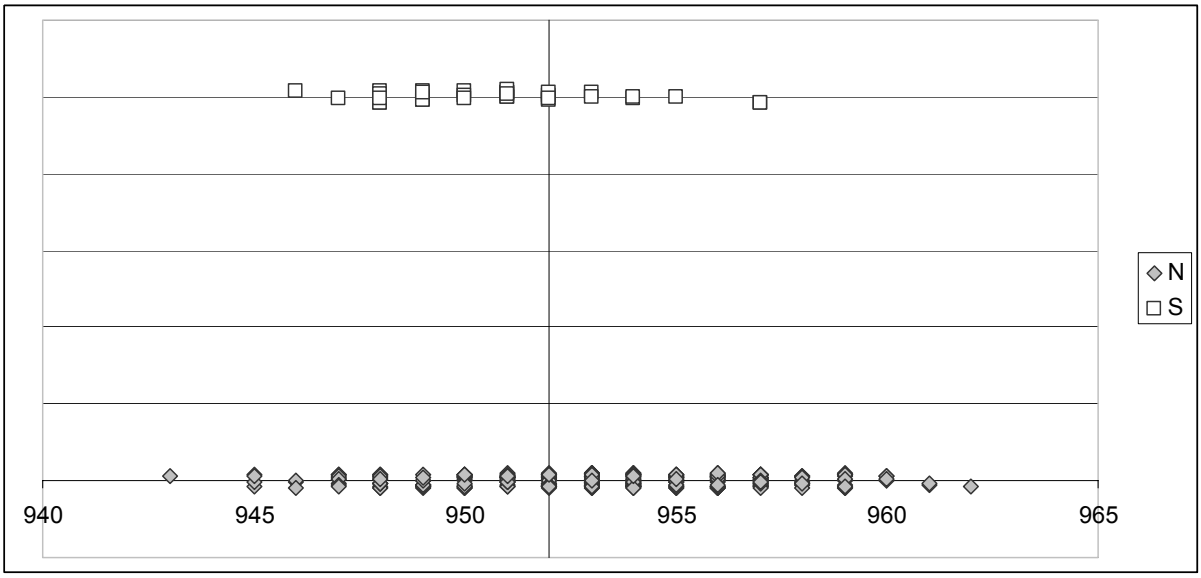

Figura 2 – Visualização da regra número 1 na Tabela 4.

Uma linha vertical marca a posição onde PressaoAtm6 = 952, permitindo a visualização da regra número 1 na Tabela 4. Os pontos plotados foram ligeiramente movidos para melhor visualização (já que muitos pontos correspondem a exatamente o mesmo valor e ocupariam a mesma posição). Nota-se na Figura 2 que não existe, para o eixo X, um ponto de corte onde possamos separar linearmente pontos com conseqüente **S** de pontos com conseqüente **N**.

A Figura 3 mostra outro exemplo da não-separabilidade linear destes dados, com uma plotagem dos dados dos atributos PressaoAtm6 e PressaoAtm0 contra o conseqüente das regras, o que corresponde à regra 5 da Tabela 4. A linha vertical corresponde ao ponto de corte PressaoAtm6 ≤ 952 e a linha horizontal ao ponto de corte PressaoAtm0 > 954. Nota-se que a área delimitada por estas duas linhas (canto superior esquerdo da figura) contém somente dados onde o conseqüente é **N**, mas pode-se observar que não existe combinação de posições destas duas linhas que separe claramente os dados com conseqüente **S** dos dados com conseqüente **N** embora seja possível imaginar conjuntos independentes de duas linhas que separem alguns registros com conseqüente **N**.

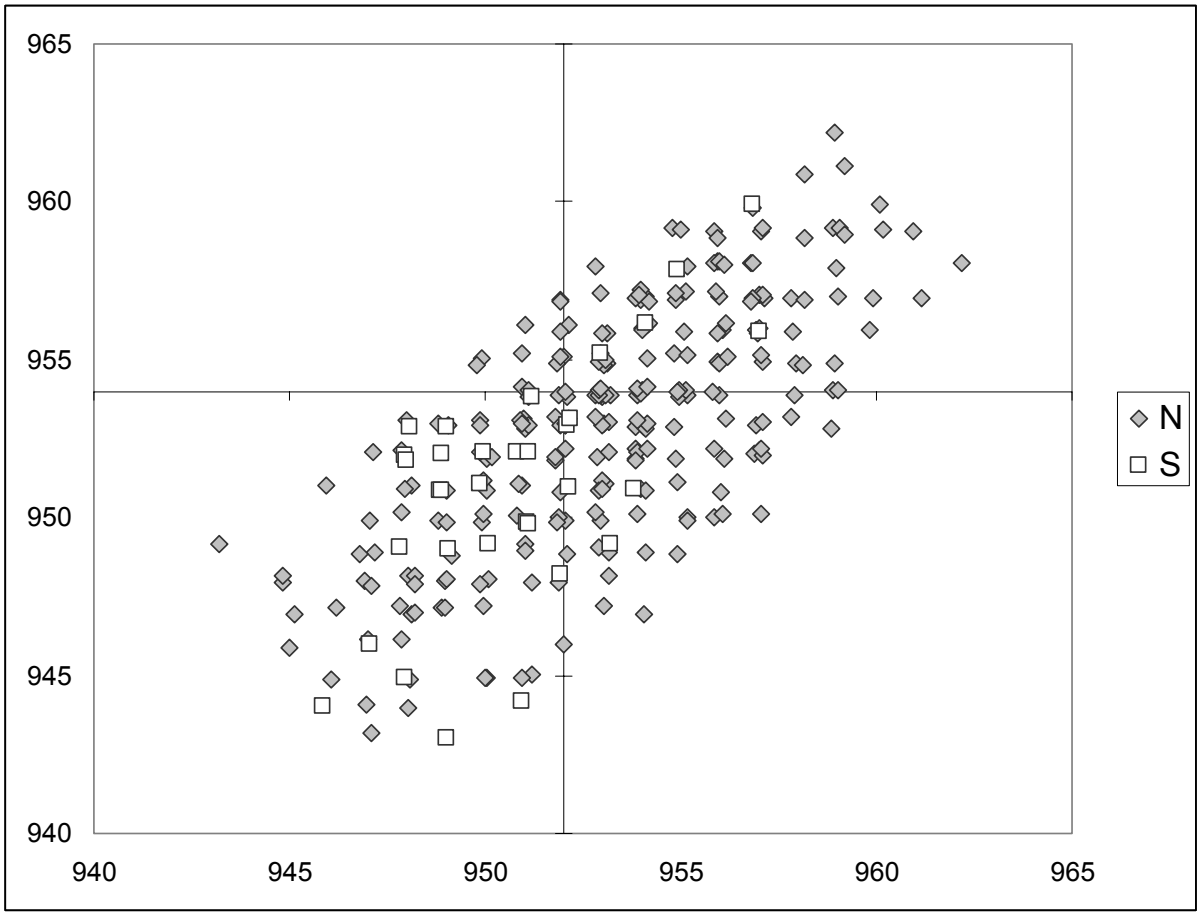

Figura 3 – Visualização da regra número 5 na Tabela 4.

Podemos observar novamente na Figura 3 que não existe uma forma simples de separar linearmente regiões retangulares no espaço de atributos de forma a ter classificações corretas para as duas classes – em outras palavras, não é possível separar linearmente as classes usando somente os atributos PressaoAtm0 e PressaoAtm6.

Outros experimentos foram executados visando investigar a possibilidade de separação linear dos dados usando vários atributos e várias combinações dos mesmos ao longo do tempo. A Tabela 5 mostra o tamanho das árvores de decisão criadas com diversas combinações de atributos entre si e tomados ao longo do tempo.

Tabela 5 – Tamanho das árvores de decisão criadas com diversas combinações de atributos.

	0	**0-3**	**0-3-6**	**0-3-6-9**	**0-3-6-9-12**
PressaoAtm	-	-	-	-	-
UmidRel	-	-	-	-	-
TempAr	-	-	-	-	-
PressaoAtm+UmidRel	-	-	-	-	29
PressaoAtm+UmidRel+TempAr	9	-	11	-	39
PressaoAtm+UmidRel+TempAr+ RadSolAcum+VelVentoMax	13	-	21	33	43
Todos	43	35	35	15	25

As linhas na Tabela 5 correspondem às combinações de atributos consideradas para os experimentos (as siglas para os atributos são mostradas na Tabela 3). A linha indicada por "Todos" contém, além de todos os atributos da Tabela 3, a direção da velocidade do máxima

do vento, a direção do vento, as temperaturas máximas e mínimas do período e a umidade interna (medida no PCD). As colunas na Tabela 5 indicam que intervalos de tempo foram considerados para coleta dos dados.

A Tabela 5 apresenta vários valores ausentes, pois para aquelas combinações não foi possível montar a árvore de decisão – o algoritmo decidiu que seria mais adequado, de acordo com seus critérios, agrupar todas as regras em uma mesma classe, às custas de erros de classificação. Os valores presentes indicam o tamanho, em nós, das árvores criadas.

A Tabela 6 mostra o percentual de dados corretamente classificados com as árvores de decisão criadas em cada experimento.

Tabela 6 – Classificação correta dos dados com as árvores de decisão criadas com diversas combinações de atributos.

	0	**0-3**	**0-3-6**	**0-3-6-9**	**0-3-6-9-12**
PressaoAtm	89,275	89,275	89,275	89,275	89,275
UmidRel	89,275	89,275	89,275	89,275	89,275
TempAr	89,275	89,275	89,275	89,275	89,275
PressaoAtm+UmidRel	89,275	89,275	89,275	89,275	94,659
PressaoAtm+UmidRel+TempAr	89,855	89,275	90,029	89,275	95,549
PressaoAtm+UmidRel+TempAr+ RadSolAcum+VelVentoMax	91,437	89,275	92,335	97,004	99,194
Todos	97,248	97,394	98,258	96,255	98,790

Os títulos das linhas e colunas da Tabela 6 são os mesmos da Tabela 5. Os valores ausentes da Tabela 5 correspondem a 89.275% de classificação correta na Tabela 6.

Apesar dos valores de classificação de alguns experimentos serem considerados altos (acima de 95%) para tarefas de classificação, os resultados são inadequados para a tarefa em vista, uma vez que é desejável a criação de uma árvore completa, e que mesmo resultados altos (89.275%) indicam que todas as regras com conseqüente **S** foram classificadas incorretamente (todas as regras foram agrupadas sem criação de árvores).

5. CONCLUSÕES E DISCUSSÃO

Neste artigo foi apresentada uma metodologia de *Data Mining* para extração de regras de uma base de dados meteorológicos e classificação das regras de acordo com uma medida de interesse. A metodologia foi aplicada à análise das regras criadas a partir de dados meteorológicos para tentar localizar regras e exceções significativas.

Alguns experimentos revelaram que com os dados considerados podemos obter regras que indicam a ausência de precipitação de acordo com registros passados, mas não foi possível obter regras simples e com boa abrangência para indicar a ocorrência de precipitação.

A complexidade da tarefa de classificação dos dados foi revelada com uma análise visual das regras obtidas, onde se observou que as classes estudadas não são linearmente separáveis. O algoritmo usado (J48, implementação do algoritmo C4.5 [4,12]), por sua natureza, não cria árvores completas, ou seja, árvores que classificam corretamente todos os dados usados para a sua criação, mesmo à custa de criar árvores complexas. Sem árvores completas não é possível avaliar que regras correspondem a *outliers* ou exceções inesperadas na base de dados. Outros algoritmos de classificação e criação de regras devem ser estudados para verifica a viabilidade

da criação de regras mais compactas e abrangentes.

O objetivo do artigo não é criar uma metodologia de *Data Mining* que permita a previsão do tempo usando somente séries temporais de alguns atributos obtidos de plataformas de coleta de dados, pois tais dados não incorporam os modelos tradicionalmente usados em previsão do tempo, mas sim demonstrar um mecanismo de análise de regras obtidas a partir destes dados para obtenção de informações de nível mais alto, isto é, descrições sumarizadas sobre estes dados.

6. REFERÊNCIAS BIBLIOGRÁFICAS

[1] S. M. Weiss e Indurkhya, N. *"Predictive Data Mining, A Practical Guide"'*, Morgan Kaufmann Publishers, Inc. San Francisco, California, 1998.

[2] K. Thearling, *"An Introduction to Data Mining"*. http://www.thearling.com. Visitado em Agosto de 2004.

[3] S. S. Haykin, *"Redes Neurais: Princípios e Prática"*. Editora Bookman, 2000.

[4] J. R. Quinlan, *"C4.5: Programs for Machine Learning"*, Morgan Kaufmann Publishers, San Mateo, CA. 1993.

[5] D. E. Goldberg, *"Genetic Algorithms in Search, Optimization and Machine Learning"*. Boston: Addison-Wesley, 1989.

[6] J. M. Keller, M. R. Gray, J. A. Givens Jr., *A Fuzzy K-Nearest Neighbor Algorithm*, in *"Fuzzy Models for Pattern Recognition: Methods that Search for Structures in Data"*, edited by J. C. Bezdek and S. K. Pal, IEEE Press, Piscataway, NJ, 258-263, 1992.

[7] D. R. Carvalho, *"Data Mining através de Indução de Regras e Algoritmos Genéticos"*. Dissertação de Mestrado em Informática Aplicada, PUCPR, PR, 1999.

[8] R. Parpinelli; H. S. Lopes; A. A. Freitas. *"Data Mining with an Ant Colony Optimization Algorithm"*. http://www.ppgia.pucpr.br/~alex/pub_papers.dir/Ant-IEEE-TEC.pdf.

[9] E. Bonabeau; M. Dorigo; G. Theraulaz. *"Swarm Intelligence, From Natural to Artificial Systems"*. Oxford University Press, 1999.

[10] B. de la Iglesia, *Mining Rules from a Database According to Multiple Measures of Interest*, in *Multiple Objective Metaheuristic Workshop*, Paris, 2002.

[11] R. C. Waltz, *"Um Estudo Climatográfico de Chuvas Máximas e Obtenção da Primeira Equação de Chuvas Intensas para São José dos Campos - SP, Tendo como Perspectiva o Planejamento Urbano do Município"*, Dissertação de Mestrado em Planejamento Urbano e Regional, Univap, 2000.

[12] I. H. Witten e E. Frank, *"Data Mining – Practical Machine Learning Tools and Techniques with Java Implementations"*, Morgan Kaufmann Publishers, Inc. San Francisco, California, 2000.