Anais do 6º Congresso Brasileiro de Agroinformática – SBIAgro 2007, São Pedro, Brasil, 8-11 outubro 2007,
Embrapa Informática Agropecuária, p.171-175.

DESCOBRINDO REGRAS DE CLASSIFICAÇÃO EM DADOS CLIMATOLÓGICOS

LAURIMAR GONÇALVES VENDRUSCULO[1]
SILVIO ROBERTO DE MEDEIROS EVANGELISTA[2]
ADRIANO FRANZONI OTAVIAN[3]
STANLEY ROBSON DE MEDEIROS OLIVEIRA[4]

RESUMO: A tradicional abordagem de descoberta do conhecimento demanda recursos financeiros e de tempo consideráveis. A mineração de dados, por meio da árvore de decisão, contribui para a geração de regras de classificação importantes. Foram analisados dados climatológicos visando à previsão do tempo e obteve-se uma árvore de classificação complexa, em profundidade e número de nós. Entretanto, o algoritmo J48, comparado com o *Decision Stump*, do ambiente WEKA, classificou corretamente 62,23% dos dados usados como treinamento.

PALAVRAS-CHAVE: mineração de dados, dados climáticos, classificadores, previsão tempo.

CLASSIFICATION RULES FROM CLIMATIC DATA

ABSTRACT: The traditional knowledge acquisition demands time and efforts considerable. Decision trees are good examples for mining data and generate classification rules. Climatic data for forecasting were analyzed using J48 and Decision Stump algorithms, from WEKA. The J48 method obtained 62,23% of data correctly classified.

KEY-WORDS: data mining, climatic data, classifier, forecasting.

1. INTRODUÇÃO

A descoberta de conhecimento, em geral, é um processo de alta complexidade e tradicionalmente demanda do especialista de domínio considerável tempo para análise acurada do problema. Além disto, requer a validação por meio de técnicas laboratoriais ou de experimentos que exigem alto custo financeiro.

Para colaborar com a automatização do processo de aquisição do conhecimento a área de aprendizado de máquina, oferece diversas técnicas que, de maneira geral, envolvem o encontro e a exploração de similaridades em conjunto de dados. Dentre os paradigmas encontrados na temática aprendizado de máquina encontra-se o de indução de regras na qual o método de árvore de decisão se encontra (LANGLEY & SIMON, 1995).

A árvore de decisão é uma estrutura de dados do tipo árvore *n*-ária, usada para a dedução de uma classe estudada (*target function*) a qual é formada adicionalmente por uma tupla de atributos (objeto a ser classificado). Segundo (BRAGA, 2004) o nó folha representa uma única classe (Ex: chuvoso), porém esta classe pode estar simbolizada em mais de um nós folhas. Um nó interno ou intermediário representa um teste sobre o valor de um atributo da tupla (Ex: Temperatura máxima > 32). E por fim, cada aresta que deixa o nó interno indo até um dos seus nós filhos representa um dos possíveis resultados do teste sobre o valor do atributo.

[1] Mestre em Engenharia Agrícola , Embrapa Informática Agropecuária – laurimar@cnptia.embrapa.br
[2] Doutor em Engenharia Elétrica, Embrapa Informática Agropecuária – silvio@cnptia.embrapa.br
[3] Bacharel em Engenharia de Computacão, Embrapa Informática Agropecuária – adriano@cnptia.embrapa.br
[4] Doutor em Ciência da Computação, Embrapa Informática Agropecuária – stanley@cnptia.embrapa.br

A técnica da árvore de decisão é aplicável quando o problema possui as seguintes características: as instâncias são representadas por tuplas de atributos com valores discretos ou contínuos; o atributo de classificação, a ser inferido na árvore retorna valores discretos; quando se tem dados errôneos no conjunto de treino ou ausência de valores.
Um dos mais populares ambientes para descoberta do conhecimento de domínio público é a ferramenta *Waikato Environment for Knowledge Analysis* – WEKA (WITTEN & FRANK, 2005). O WEKA fornece uma API para suporte ao aprendizado de máquina, implementado em Java, que incorpora vários algoritmos para: seleção de atributos, seleção de instâncias, algoritmos de aprendizagem supervisionada e outros. Esta ferramenta possui vários classificadores dentre estes utilizamos neste estudo o J48, baseado no algoritmo Quinlan's C4.5.
Este trabalho tem como motivação a complexidade dos fenômenos atmosféricos envolvidos no processo de previsão do tempo. A moderna previsão do tempo está associada atualmente a utilização de modelos matemáticos para descrever o estado futuro da atmosfera. Existem duas abordagens para este assunto: a abordagem determinística, ou seja dado uma condição inicial existe apenas um sistema final de previsão do tempo. A outra abordagem, chamada previsão por conjuntos visa aperfeiçoar as limitações da primeira que não contempla o comportamento caótico da atmosférica e a não representação da totalidade dos fenômenos atmosféricos. A previsão por conjuntos consiste em utilizar diversas previsões, ou seja por meio de condições iniciais ligeiramente diferentes pode-se obter modelos ligeiramente modificados (Portal Previsões Numéricas, 2007).
O presente estudo propõe o estudo de classificadores supervisionados com alto percentual de acurácia, que contribuam para a tomada de decisão na área de previsão do tempo.

2 – MATERIAL E MÉTODOS

O estudo proposto utilizou a ferramenta WEKA, os algoritmos J48 e *Decision Stump* que implementam a árvore de decisão.
O método J48 induz a produção de regras (IF THEN ELESE) por meio de um conjunto de dados de treinamento. Este conjunto de dados pode conter atributos discretos ou contínuos. O modelo da árvore de decisão no J48, é construído em função do grau de aleatoriedade dos valores que um atributo *A* pode assumir. Esta aleatoriedade é verificada em função do cálculo de entropia dos dados, representada pela equação 1. A entropia para o atributo *A* H(A) é dada por:

$$H(A) = -\sum_{v=1}^{m} P_V \, LOG_2 \, P_V \quad (1)$$

onde *m* são os valores que podem ser assumidos por *A* e *p* é a probabilidade de *A* ser igual ao valor cujo índice é *v* .O J48 também calcula o ganho de informação. Este parâmetro está diretamente relacionado à entropia condicional.

O método *Decision Stump* gera uma árvore de decisão de apenas um nível onde o atributo eleito para formar esta árvore é obtido por meio do ganho de informação (entropia). Apesar de sua aparente simplicidade, o método alcança bons resultados e pode ser utilizado para atributos numéricos, simbólicos ou classes que contenham ambos os tipos.
Os testes foram realizados em um computador AMD Semprom, com 1 MB de memória RAM, na plataforma Windows 2000. Utilizou-se a versão 3.4.10 do software WEKA

Com esta configuração não foi possível executar o classificador *Random Forest*, o qual exigiu maior quantidade de memória RAM do aquela disponível.
Os dados utilizados neste estudo são constituído por um conjunto de dados, divididos em 66% para treino e 34% para teste. O conjunto de dados possui 132.627 instâncias, com cinco atributos, incluindo a classe nomeada de CLASSTEMPO como mostrado a Figura 1. A base de dados foi obtida por meio de uma consulta SQL à base de dados do Sistema de Monitoramento Agrometeorológico Agritempo, gerenciada pelo SGBD Oracle, versão 8.1.
O sistema Agritempo[5] recebe diariamente, de seus parceiros institucionais, informações de estações climatológicas de todo o território brasileiro. São gerados produtos na forma de mapas de monitoramento e previsão bem como os dados fornecem subsídios para a elaboração de boletins agrometeorológicos e zoneamento agrícola.
Os dados analisados provém da tabela de previsão do tempo e refletem o período de dados diários de 01/01/2005 a 05/05/2007.

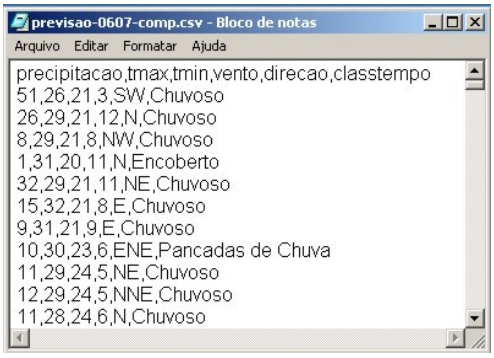

Figura 1 - Porção inicial dos dados analisados.

3 – DISCUSSÃO DE RESULTADOS

Ao utilizar o classificador J48, as nove classes estabelecidas (Chuvoso, Encoberto, Céu Claro, Pancadas de Chuva, Nublado com Chuva, Nublado, Parcialmente Nublado, Pancadas Isoladas),o volume de instâncias e a variação de valores dos atributos numéricos resultaram em uma árvore de decisão altamente complexa. Esta situação não permitiu a construção de regras simples, apesar da rapidez da classificação dos dados de treinamento (~ 8 minutos).
O decréscimo no valor do fator de confiança no J48 resultaram em árvores com menores números de nós, conforme indica a Tabela 1. Nota-se que a redução do tamanho da árvore é proporcional ao fator de confiança, porém sem causar impacto na taxa de acerto.

Tabela 1 – Resultados da aplicação do J48 para fatores de confiança distintos

Fator de confiança	Número nós	Taxa de acerto (%)
0,25	10.818	62,23
0,15	7.400	62,07
0,05	3.644	61.17

A regra, apresentada a seguir como exemplo, especifica acuradamente os parâmetros vento, temperatura máxima, direção e precipitação, os quais representam os fatores de maior importância para a classificação do tempo como chuvoso.

IF (“2 > Vento <= 12”) **AND** “Tmax <= 22” **AND** “direcao = NE” **AND** (“5 > precipitacao > 9”) THEN Chuvoso (11.0/2.0)

[5] Disponível na URL : http://www.agritempo.gov.br

O algoritmo *Decision Stump* foi usado na mesma base de dados para fins de comparação de performance e a taxa de acerto alcançou 45,7 %. Neste método, a árvore de decisão é simplificada a um único nível e o atributo eleito é foi a precipitação. As regras encontradas por este método, mostradas a seguir, são genéricas e devem ser usadas apenas para indicar o atributo que melhor explica as classes relativas às condições climatológicas.

precipitação <= 0.5 ⇒ Parcialmente Nublado
precipitação > 0.5 ⇒ Pancadas de Chuva
precipitação onde o valor for desconhecido ⇒: Pancadas de Chuva

4 - CONCLUSÕES

O classificador J48, disponível no aplicativo WEKA mostrou um desempenho melhor que o método *Decison Stump*. com relação a análise dos dados.
A árvore de decisão do J48 encontrada neste estudo, apesar da maior taxa de acerto, ainda não pode ser utilizada como substituto aos modelos matemáticos. Estes modelos determinísticos ainda são superiores, dado a alto grau de incerteza nesta temática. Todavia, os classificadores são instrumentos complementares importantes na temática da agrometeorologia, considerada a rapidez e acurácia de seus resultados.
O conjunto de dados climatológicos do Agritempo mostrou valioso para os testes dos métodos de classificação estudados, em função de seu volume e veracidade. Este conjunto pode tornar-se útil quando se avalia a robustez e desempenho de novos métodos de mineração de dados. No estudo em questão , prevê-se em trabalhos futuros ampliar a janela temporal estudada, pois em agroclimatologia o período aconselhável para predições varia de 20 a 30 anos.
Este trabalho mostrou que os classificadores são ferramentas importantes para a predição das condições do tempo, entretanto existem ainda grandes desafios. Novas metodologias precisam considerar os aspectos de temporalidade e espacialidade dos dados. Por exemplo, fenômenos localizados, como chuvas de convecção (chuvas de verão) contribuem para predizer erroneamente as condições de grandes áreas. Os esforços futuros serão focados na investigação de tendência em grandes períodos e da ocorrência de variações sazonais ou cíclicas dos dados.

5 – REFERÊNCIAS BIBLIOGRÁFICAS

BRAGA, B. da R. ; Jr. D´ALMEIDA, J. N.; BAIÃO, F; MATTOSO, M.L.Q. **IDSMiner: Data Mining de Modelos de Detecção de Intrução – Relatório Técnico do Projeto ClusteMiner – Janeiro/2004** Disponível em: < http://clusterminer.nacad.ufrj.br/TechReport/RT06.pdf>. Consultado em: 3 de maio de 2007.

LANGLEY, P.; SIMON, H. A Applications of machine learning and rule induction. **Communications of de ACM**, , v. 38, n. 11 p. 54-64, November 1995

PORTAL PREVISÕES NUMÉRICAS, O sistema de Previsão de tempo global por ensemble do CPTEC. Disponível : < http://www.cptec.inpe.br/prevnum/exp_ensemble.shtml> , Consultado em :2 maio de 2007.

WITTEN, I.H.; FRANK, E. **Data Mining: Practical machine learning tools and techniques**, Morgan Kaufmann, San Francisco, 525 p.,2005