9ª MOSTRA ACADÊMICA UNIMEP
08 a 10/11 de 2011
Tema: "Ambiente e Sustentabilidade"

19 Congresso de Iniciação Científica

INCLUSÃO DE UM NOVO ALGORITMO DE CLASSIFICAÇÃO NA FERRAMENTA KIRA

Autor(es)

MIRELA TEIXEIRA CAZZOLATO

Orientador(es)

MARINA TERESA PIRES VIEIRA

Apoio Financeiro

PIBIC/CNPQ

1. Introdução

Grandes quantidades de dados são coletados e armazenados diariamente nas instituições. O Processo de KDD (*Knowledge Discovery in Database*) é uma maneira automatizada de auxiliar na exploração do conteúdo desses dados, buscando extrair informações potencialmente úteis.

De acordo com Fayyad, Piatetsky-Shapiro e Smith (1996), a descoberta de conhecimento em bases de dados consiste em um processo não trivial de identificação de padrões contidos nos dados que sejam válidos, novos, potencialmente úteis e compreensíveis.

Conforme Han e Kamber (2006) trata-se de um processo composto por sete etapas: limpeza, integração, seleção, transformação, mineração, avaliação dos padrões e apresentação do conhecimento.

A principal etapa do processo de KDD é a mineração de dados (*Data Mining*) e consiste, segundo Witten, Frank e Hall (2011), na extração de informações implícitas, previamente desconhecidas e potencialmente úteis dos dados. É nessa etapa que os algoritmos de mineração são aplicados. O conhecimento obtido pode ser aplicado em diversas situações, tais como tomada de decisões e controle de processo.

Existem diversas tarefas de mineração; algumas das principais são: **classificação**, **regras de associação**, **e clusterização**. Os assuntos tratados neste artigo têm como foco a tarefa de classificação, por ser a tarefa utilizada durante o projeto de pesquisa.

Segundo Han e Kamber (2006), a classificação consiste em um processo de descoberta de um modelo (ou função) que descreve e distingue classes de dados ou conceitos. Esse modelo é obtido por meio da análise dos dados de treinamento, cujo rótulo da classe é conhecido, e é utilizado para prever a classe de outros objetos.

Durante o processo de classificação é utilizado o **atributo classe**, que indica, para cada instância, sua classe correspondente. Ele pode ser escolhido no conjunto de dados criado ou a partir de outros atributos.

Conforme Han e Kamber (2006), o processo de classificação é dividido em duas fases, a de treinamento e a de classificação:

- **Fase de treinamento:** é construído um classificador que descreve um conjunto de classes, podendo ser utilizado para classificar futuras tuplas de dados;
- **Fase de classificação:** analisada sua precisão, o modelo construído pode ser usado na classificação de tuplas de dados cujo rótulo é desconhecido.

Existem vários algoritmos de classificação, tais como o ID3 (QUINLAN, 1986), o C4.5 (QUINLAN, 1993) e o DTree (BORGELT, 2003).

Um importante ponto a se considerar é a apresentação do modelo construído. Ele pode ser representado de diversas formas, tais como por meio de regras de classificação SE-ENTÃO e árvores de decisão.

De acordo com Han e Kamber (2006), **árvore de decisão** é uma representação gráfica intuitiva e fácil de assimilar das regras de classificação. Ela é composta por: **nó** (representa um teste em um valor de atributo), **ramo** (representa um resultado de teste) e **folhas** (representam classes ou distribuições de classe).

Após a criação do modelo de classificação, é necessário estimar a precisão do classificador. Existem diversas formas de avaliar a precisão do classificador, tais como: **matriz de confusão**, que fornece uma visão geral do resultado da classificação; **estatística Kappa**, utilizada para medir o sucesso da predição; **ROC Area**, que informa a precisão das classificações para cada classe; dentre outras.

Um exemplo de ferramenta de mineração de dados de código aberto é a WEKA. Conforme Witten, Frank e Hall (2011), a WEKA (Waikato Environment for Knowledge Analysis) é uma coleção de algoritmos e ferramentas de processamento de dados desenvolvida na Universidade de Waikato, na Nova Zelândia.

Para o uso da WEKA e de outras ferramentas existentes é preciso conhecer o processo que envolve a mineração de dados, que não é trivial, pois essas ferramentas não oferecem recursos para guiar o usuário a realizar esse processo. Procurando oferecer facilidades para auxiliar a realizar o processo de mineração de dados, foi desenvolvida a ferramenta Kira, por Mendes (2009). Essa ferramenta tem como objetivo ensinar o usuário, com pouco conhecimento sobre o processo envolvido na mineração, a preparar os dados, escolher a tarefa de mineração adequada e analisar os resultados obtidos, abstraindo boa parte do conhecimento exigido.

Sua arquitetura é composta por três módulos:

Sua arquitetura é composta por três módulos:

• **Módulo de apoio à origem:** são identificadas as fontes de dados que serão utilizadas;

• **Módulo de apoio à preparação:** são executadas as atividades de preparação de dados, contemplando a integração, limpeza, seleção e transformação dos dados.

• **Módulo de apoio à análise:** são executadas as etapas referentes à mineração de dados e análise dos resultados obtidos. São exibidos os dados selecionados e transformados e executado o algoritmo minerador, mostrando em seguida os resultados obtidos, auxiliando o usuário em sua análise.

Em cada etapa são fornecidos guias que auxiliam o usuário a executar passo a passo o processo envolvido na mineração.

Conforme Vieira et al. (2009), com o uso da ferramenta Kira em salas de aula, observou-se que alunos têm comprovado o potencial da ferramenta para o processo de ensino/aprendizagem de mineração de dados. Experimentos realizados mostraram que é mais fácil ensinar o processo envolvido na mineração de dados utilizando a ferramenta Kira do que utilizando outra ferramenta de mineração de dados.

No trabalho de Mendes (2009) foram desenvolvidas funcionalidades para apoiar a mineração de dados utilizando Regras de Associação. Em (CAZZOLATO, 2010) foi incorporado o módulo de classificação à ferramenta, disponibilizando guias e um algoritmo para auxiliar a tarefa de Classificação, com base no trabalho de Coelho (2010).

2. Objetivos

O trabalho aqui apresentado teve como objetivo dar continuidade no módulo de classificação da ferramenta Kira, visando a inclusão de um novo algoritmo e de novos recursos para a tarefa de classificação. Objetivou-se com isso tornar a ferramenta mais abrangente com relação ao ensino de mineração de dados usando a tarefa de classificação.

3. Desenvolvimento

Com base em testes realizados com diferentes algoritmos de classificação presentes na ferramenta WEKA, foi escolhido o algoritmo J48 para a inclusão na ferramenta Kira.

Na ferramenta Weka, o algoritmo J48 é uma implementação do algoritmo C4.5. Conforme Witten, Frank e Hall (2011), a versão do algoritmo J48 presente na ferramenta implementa a revisão 8 do algoritmo C4.5; trata-se de uma versão ligeiramente melhorada, sendo a última pública desta família de algoritmos, antes que a implementação comercial C5.0 fosse lançada.

Segundo Rokach e Maimon (2008), o C4.5 é uma evolução do ID3, apresentada pelo mesmo autor Ross Quinlan em (QUINLAN, 1993); esse algoritmo, ao contrário do ID3, lida com atributos numéricos e também pode induzir a formação de um conjunto que incorpora os valores faltantes.

Foi incorporado o algoritmo J48 na ferramenta Kira, e com base em estudos e experimentos sobre a tarefa de classificação, foram elaboradas interfaces para auxiliar o usuário no processo de aprendizagem da mineração de dados utilizando a tarefa de classificação, fornecendo também explicações dos resultados obtidos pelo algoritmo.

Por fim, foram realizados testes com diferentes conjuntos de dados e efetuados ajustes. Todo o processo de incorporação da classificação foi realizado com o cuidado de desenvolver interfaces de fácil assimilação por parte do usuário.

4. Resultado e Discussão

Como citado anteriormente, neste projeto foi incorporado o algoritmo J48 no módulo de classificação na Kira, tendo como objetivo disponibilizar funcionalidades que auxiliassem no processo de aprendizagem da mineração de dados utilizando a tarefa de classificação. Além disso, foi desenvolvido um estudo de caso para a aplicação da tarefa de classificação na ferramenta Kira.

O estudo de caso desenvolvido diz respeito a um conjunto de dados de uma empresa financeira, em que se deseja estimar o risco de se conceder um empréstimo a um cliente com base em seu perfil.

Para a execução da mineração de dados na ferramenta Kira, o usuário deve primeiramente carregar os dados com os quais deseja trabalhar, selecionar a tarefa a ser utilizada, definir o problema a ser resolvido, o objetivo a ser alcançado e selecionar os dados a serem utilizados.

Essas funcionalidades são as mesmas dos algoritmos já disponíveis na ferramenta. Feito isso, a próxima etapa a ser realizada é a Seleção dos Dados. Desse ponto do processo de mineração de dados em diante foram desenvolvidos os recursos para a tarefa de classificação utilizando o J48.

Depois de selecionar os dados a serem utilizados e indicar o atributo classe, o usuário deve escolher o algoritmo de classificação a ser utilizado. Feito isso, ocorre a etapa de mineração de dados, em que o algoritmo é executado. São disponibilizadas então formas de visualização e avaliação dos padrões obtidos.

Ao executar o algoritmo J48, são apresentadas como visualização e avaliação dos resultados a árvore de decisão, estatísticas e medidas de acurácia do classificador. Para dar suporte a esses resultados foram desenvolvidas interfaces, sempre tomando o devido cuidado de torná-las o mais simples e intuitivas possível.

A árvore de decisão gerada pelo algoritmo J48 é mostrada na Figura 1. Para exemplificar, o conjunto de dados utilizado nas interfaces presentes neste artigo é o mesmo do estudo de caso referente a uma empresa finaceira, desenvolvido durante o projeto de pesquisa.

Ao apresentar as estatísticas geradas pelo algoritmo, procurou-se incorporar uma explicação do significado de cada medida apresentada, como pode ser visto no exemplo apresentado na Figura 2 em que é mostrada a explicação da estatística Kappa. Essa informação é mostrada na parte inferior da interface da ferramenta. Para incorporar essas explicações foi necessário estudar os significados das várias estatísticas fornecidas pelo classificador (estatística Kappa, erro absoluto relativo, etc.). Na Figura 2 é mostrada a explicação.

Para a análise da acurácia do classificador são disponibilizadas medidas como Precisão e F-Measure, além da matriz de confusão, como pode ser visto na Figura 3. Também foi necessário realizar estudos sobre essas medidas estatisticas, para poder disponibilizar uma descrição ao usuário, de modo a auxiliar no seu entendimento. Na Figura 3, ao selecionar um dos valores (partes a), sua explicação é exibida na parte inferior de cada tabela (partes b).

As explicações das estatísticas foram elaboradas por meio de estudos utilizando diversos materiais bibliográficos e também com o suporte da professora da Universidade Metodista de Piracicaba, Maria Imaculada de Lima Montebello, que atua na área de Estatística.

De forma geral, as atividades desenvolvidas durante o projeto foram:

- Melhorias no módulo de classificação desenvolvido em (CAZZOLATO, 2010);
- Participação na elaboração de um artigo que apresenta os recursos existentes na ferramenta Kira para apoiar a classificação, que foi submetido à revista RITA;
- Avaliação da ferramenta Kira em sala de aula com alunos do curso de mestrado em Ciência da Computação da Unimep;
- Experimentos e escolha de um dos algoritmos da ferramenta WEKA para a incorporação na ferramenta Kira;
- Especificação do módulo de classificação para o algoritmo J48;
- Inclusão do algoritmo J48 na ferramenta Kira, procurando incorporar instruções e recursos na interface para auxiliar a aplicação da tarefa de classificação;
- Elaboração de um estudo de caso para a tarefa de classificação;
- Testes utilizando diferentes conjuntos de dados e testes com usuários para reconhecer pontos a serem refinados na ferramenta.

A inclusão de outro algoritmo de classificação na ferramenta Kira incorporou mais recursos para auxiliar o ensino de mineração de dados usando a tarefa de classificação.

5. Considerações Finais

Durante a revisão bibliográfica obteve-se um conhecimento teórico mais abrangente sobre o processo envolvido na mineração de dados utilizando a tarefa de classificação.

Na incorporação do algoritmo J48 na ferramenta Kira, obteve-se um bom conhecimento na utilização da linguagem de programação Java. Com base em testes e discussões em grupo, foram efetuadas modificações no módulo incorporado, buscando a melhor forma de guiar o usuário na execução do processo envolvido na mineração de dados, utilizando a tarefa de classificação.

O algoritmo de classificação já foi totalmente incorporado na ferramenta e está funcionando corretamente. Como próximas etapas pretende-se realizar testes com mais usuários, de modo a identificar necessidades de refinamentos nas instruções que orientam a execução do módulo de classificação.

A utilização de dois algoritmos de classificação possibilita que usuários o utilizem sobre um mesmo problema, podendo comparar os resultados de cada um, identificar o comportamento deles diante do problema específico, para escolha do mais adequado.

Referências Bibliográficas

BORGELT, C.. ***Decision and Regression Trees - dti/dtp/dtx/dtr/rsx - induce, prune, and execute decision and regression trees.*** Disponível em: http://www.borgelt.net/doc/dtree/dtree.html. Acesso em: 10 ago. 2010. Copyright © 1996-2003 Christian Borgelt, 2003.

CAZZOLATO, M. T. **Incorporação da Tarefa de Classificação na Ferramenta de Mineração de Dados Kira.** Projeto de Iniciação Científica, Programa PIBIC/CNPq. Universidade Metodista de Piracicaba. Protocolo CONSEPE nro 12909, Ago/2009-Jul/2010.

COELHO, U. M.. **Ferramenta Instrucional para Mineração de Dados Usando Classificação.** 2010. Dissertação (Programa de Mestrado em Ciência da Computação) – Faculdade de Ciências Exatas e da Natureza, Universidade Metodista de Piracicaba, Piracicaba, 2010.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMITH, P. ***From Data Mining to Knowledge Discovery: An Overview.*** In: Advances in Knowledge Discovery and Data Mining, AAAI Press/ The MIT Press, MIT, Cambridge, Massachusets, England, 1996.

HAN, J.; KAMBER, M. ***Data Mining - Concepts and Techniques.*** 2a edição. Nova York: Morgan Kaufmann, 2006.

MENDES, Eduardo Fernando. **Automatização da técnica de mineração de dados auxiliada por guias.** 2009. 115 f. Dissertação (Programa de Mestrado em Ciência da Computação) – Faculdade de Ciências Exatas e da Natureza, Universidade Metodista de Piracicaba, Piracicaba, 2009.

QUINLAN, J.R. ***Induction of Decision Trees.*** Kluwer Academic Publishers. Boston, 1986.

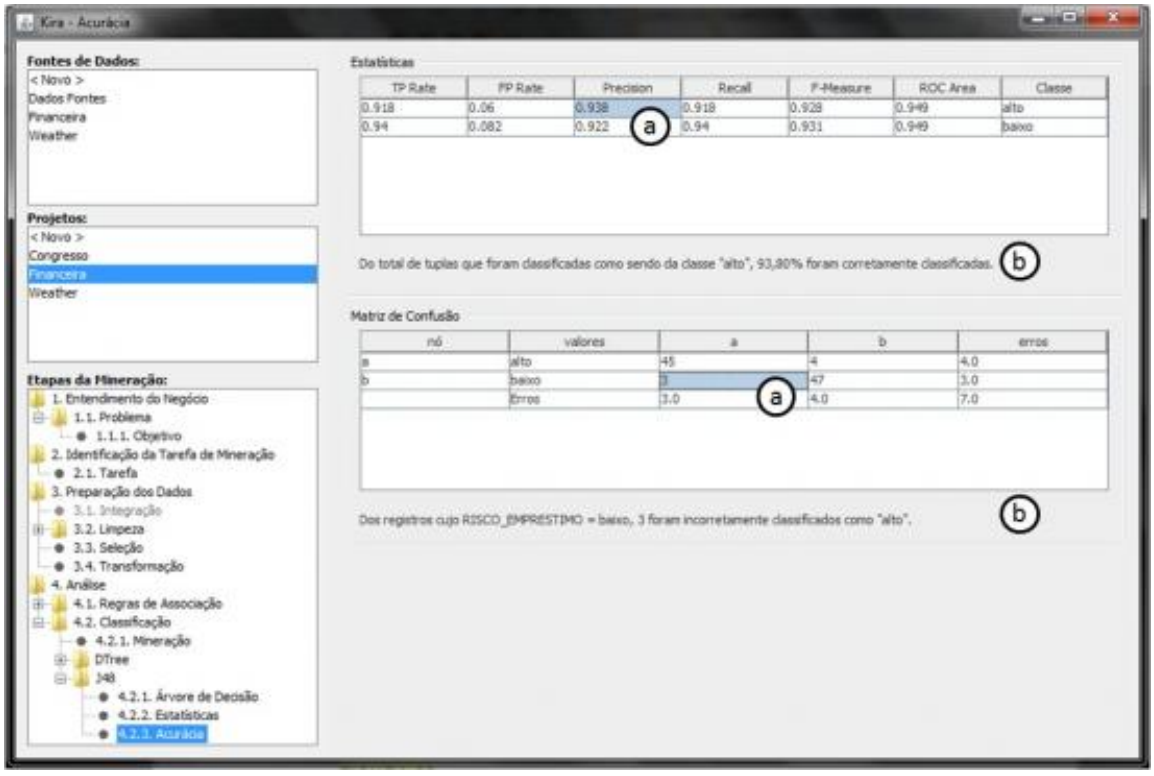
QUINLAN, J.R. C4.5: ***Programs for Machine Learning.*** São Francisco: Morgan Kaufmann, 1993.

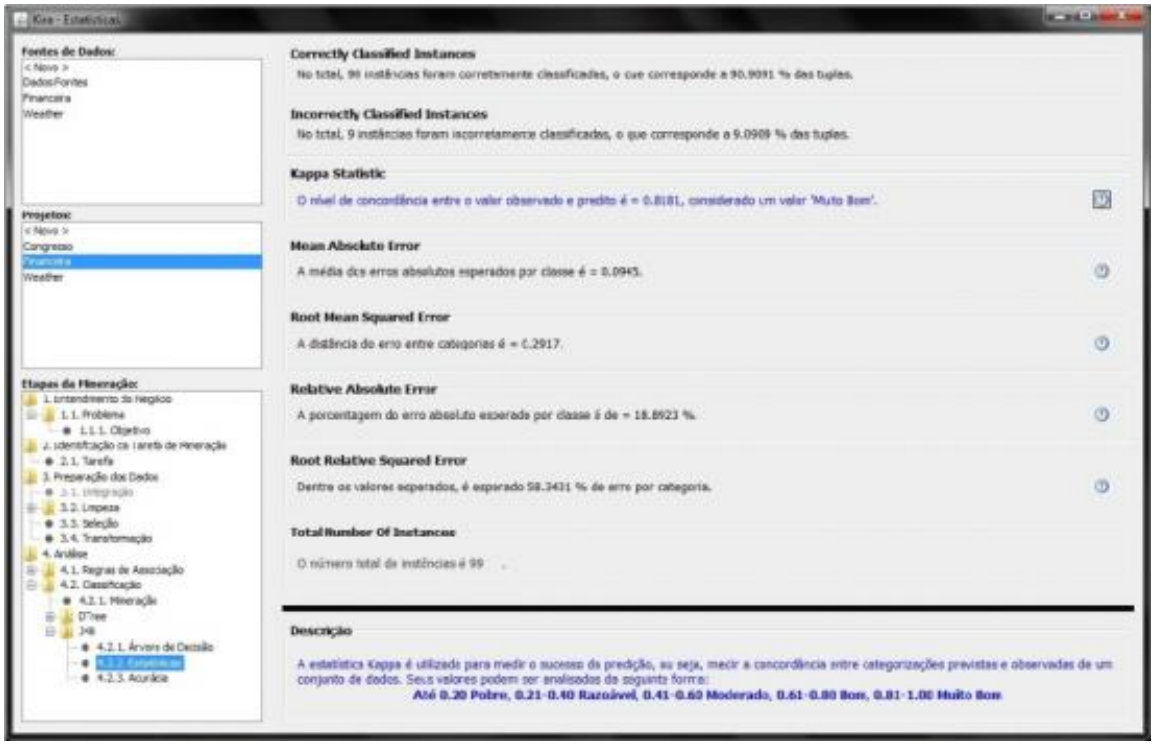
ROKACH, L., MAIMON, O.. ***Data Mining with Decision Trees - Theory and Applications. Series in Machine Perception and Artificial Intelligence*** - Vol. 69; World Scientific Publishing Co. Pte. Ltd. 2008.

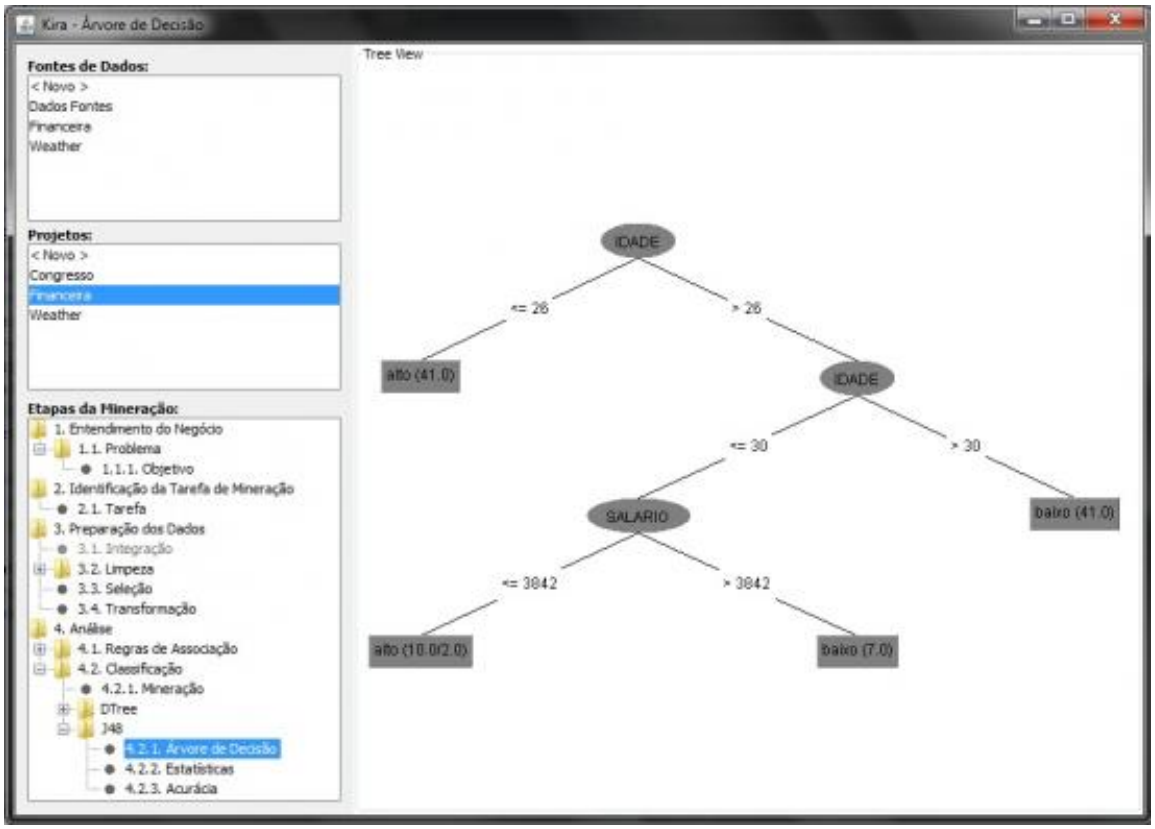
VIEIRA, M. T. P.; SILVA, A. E. A.; PEIXOTO, C.S.A.; MENDES, E. F.; GOMIDE, R. S.. Kira – ***A Tool Based On Guides And Domain Knowledge To Instruct Data Mining.*** In: IADIS International Conference Applied Computing, 2009, Roma. v. II. p. 12-16.

WITTEN, I. H.; FRANK, E.; HALL, M. A. ***Data Mining - Pratical Machine Learning Tools and Techinques.*** 3a Edição. Massachusetts, USA. 2011.

Anexos




Kira - Estatísticas
Fontes de Dados:
< Novo >
Dados Fontes
Financeira
Weather
Projetos:
< Novo >
Congresso
Financeira
Weather
Etapas da Mineração:
1. Entendimento do Negócio
1.1. Problema
1.1.1. Objetivo
2. Identificação da Tarefa de Mineração
2.1. Tarefa
3. Preparação dos Dados
3.1. Integração
3.2. Limpeza
3.3. Seleção
3.4. Transformação
4. Análise
4.1. Regras de Associação
4.2. Classificação
4.2.1. Mineração
DTree
J48
4.2.1. Árvore de Decisão
4.2.2. Estatísticas
4.2.3. Acurácia
Correctly Classified Instances
No total, 90 instâncias foram corretamente classificadas, o que corresponde a 90.9091 % das tuplas.
Incorrectly Classified Instances
No total, 9 instâncias foram incorretamente classificadas, o que corresponde a 9.0909 % das tuplas.
Kappa Statistic
O nível de concordância entre o valor observado e predito é = 0.8181, considerado um valor 'Muito Bom'.
Mean Absolute Error
A média dos erros absolutos esperados por classe é = 0.0945.
Root Mean Squared Error
A distância do erro entre categorias é = 0.2917.
Relative Absolute Error
A porcentagem do erro absoluto esperada por classe é de = 18.8923 %.
Root Relative Squared Error
Dentre os valores esperados, é esperado 58.3431 % de erro por categoria.
Total Number Of Instances
O número total de instâncias é 99
Descrição
A estatística Kappa é utilizada para medir o sucesso da predição, ou seja, mede a concordância entre categorizações previstas e observadas de um conjunto de dados. Seus valores podem ser analisados da seguinte forma:
Até 0.20 Pobre, 0.21-0.40 Razoável, 0.41-0.60 Moderado, 0.61-0.80 Bom, 0.81-1.00 Muito Bom


Kira - Árvore de Decisão
Fontes de Dados:
< Novo >
Dados Fontes
Financeira
Weather
Projetos:
< Novo >
Congresso
Financeira
Weather
Etapas da Mineração:
1. Entendimento do Negócio
1.1. Problema
1.1.1. Objetivo
2. Identificação da Tarefa de Mineração
2.1. Tarefa
3. Preparação dos Dados
3.1. Integração
3.2. Limpeza
3.3. Seleção
3.4. Transformação
4. Análise
4.1. Regras de Associação
4.2. Classificação
4.2.1. Mineração
DTree
J48
4.2.1. Árvore de Decisão
4.2.2. Estatísticas
4.2.3. Acurácia
Tree View
IDADE
<= 26
> 26
alto (41.0)
IDADE
<= 30
> 30
SALARIO
baixo (41.0)
<= 3842
> 3842
alto (10.0/2.0)
baixo (7.0)