8ª MOSTRA ACADÊMICA UNIMEP
26 a 28/10 de 2010
Tema: Desafios da Educação Superior na Agenda do Novo Milênio

18º Congresso de Iniciação Científica

INCORPORAÇÃO DA TAREFA DE CLASSIFICAÇÃO NA FERRAMENTA DE MINERAÇÃO DE DADOS KIRA

Autor(es)

MIRELA TEIXEIRA CAZZOLATO

Orientador(es)

MARINA TERESA PIRES VIEIRA

Apoio Financeiro

PIBIC/CNPQ

1. Introdução

Diariamente, grandes empresas como bancos e supermercados geram uma grande quantidade de dados. A coleta e o armazenamento desses dados têm tomado dimensões cada vez maiores. Devido ao seu grande volume, analisar esses dados, sem auxilio de ferramentas computacionais, é cada vez mais inviável.

Também conhecido como o Processo de KDD (*Knowledge Discovery in Databases*), o Processo de Descoberta de Conhecimento em Bases de Dados surge como uma maneira automatizada de auxiliar a exploração do conteúdo de grandes quantidades de dados.

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

De acordo com Han e Kamber (2006), trata-se de um processo dividido em sete etapas: **limpeza**, **integração**, **seleção**, **transformação** , **mineração**, **avaliação dos padrões** e **apresentação do conhecimento**.

A mineração de dados (*Data Mining*) é considerada a mais importante etapa do processo de KDD. Nela são aplicados algoritmos que buscam extrair conhecimento contido em grandes quantidades de dados, buscando identificar padrões que explicitam ocorrências nos dados, não observadas anteriormente.

Conforme Han e Kamber (2006), a mineração de dados é um campo multidisciplinar que envolve diversas áreas, tais como tecnologia de banco de dados, estatística, redes neurais e inteligência artificial.

O conhecimento obtido pode ser aplicado em situações como tomada de decisões, controle de processo e processamento de consultas.

Existem diversas tarefas de mineração. Algumas das principais são: r**egras de associação**, **classificaçã**o e **clusterização**.

Os assuntos tratados neste artigo focam a tarefa de classificação. De acordo com Han e Kamber (2006), classificação é o processo de descoberta de um modelo (ou função) que descreve e distingue classes de dados ou conceitos. Isso é realizado com o propósito de usar o modelo para prever a classe de objetos, cujos rótulos são desconhecidos. O modelo é obtido através da análise de dados de treinamento, ou seja, através de objetos de dados cujo rótulo da classe é conhecido.

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

Conforme Han e Kamber (2006), a classificação é um processo dividido em dois passos:

• **Fase de treinamento:** é construído um classificador que descreve um conjunto predeterminado de classes de dados ou conceitos. As regras geradas podem ser utilizadas para classificar futuras tuplas de dados e proporcionar uma visão mais detalhada do conteúdo do banco de dados.

• **Fase de classificação:** se a precisão do classificador for considerada aceitável, o modelo construído pode ser usado para classificar futuras tuplas de dados, cujo rótulo é desconhecido.

Conforme Tan, Steinbach e Kumar (2006), um modelo de classificação é útil para as finalidades:

• **Modelagem descritiva:** o modelo serve como uma ferramenta explicativa dos dados analisados.

• **Modelagem preditiva:** o modelo é usado para prever o rótulo de classes de registros desconhecidos; o modelo é tratado como uma "caixa preta", atribuindo automaticamente um rótulo de classe quando apresentado um conjunto de atributos de um registro desconhecido.

O modo de apresentação do modelo obtido é um importante ponto a considerar. Ele pode ser representado de várias formas como: **regras de classificação SE-ENTÃO**, **árvores de decisão**, **fórmulas matemáticas** e **redes neurais**.

De acordo com Han e Kamber (2006), **árvore de decisão** é uma representação gráfica das regras de classificação. Semelhante à estrutura de árvore, uma árvore de decisão 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).

Ainda segundo Han e Kamber (2006), as árvores de decisão podem manipular dados de alta dimensão. A representação do conhecimento adquirido em forma de árvore é geralmente intuitiva e fácil de assimilar. Como qualquer outro método, o uso bem sucedido de árvores de decisão pode depender dos dados disponíveis.

Um ponto relevante na representação por árvore de decisão é a facilidade de sua conversão para regras de classificação.

Conforme Witten e Frank (2005), em sua representação, considerando uma regra do tipo “SE condição ENTÃO conclusão”, a **condição** de uma regra é uma série de testes, como os testes em nós feitos em árvores de decisão; a **conclusão** fornece a classe (ou classes) que se aplicam aos casos abrangidos por esta regra. Além de serem facilmente geradas, essas regras são de fácil interpretação por parte do usuário.

Após a criação do modelo de classificação, é estimada a precisão do classificador. Para isso, utilizam-se tuplas de dados de teste, antes da realização da fase de classificação, onde o modelo construído é utilizado para a classificação.

A avaliação de desempenho de um modelo de classificação é baseada no número de registros de teste que foram corretamente e incorretamente preditos. Uma forma de representação conhecida é a chamada matriz de confusão.

A **matriz de confusão** fornece uma visão geral do resultado da classificação. Após a construção do modelo de classificação e sua aplicação, cabe ao usuário analisar se ele é adequado. Essa análise deve ser baseada nos erros e acertos cometidos pelo classificador.

É comum o uso de ferramentas que auxiliem na realização do Processo de KDD. Atualmente existem diversas ferramentas disponíveis para a execução de diferentes tarefas de mineração, mas elas exigem do usuário um grande conhecimento sobre o processo de descoberta de conhecimento.

A ferramenta Kira tem como objetivo ensinar como preparar os dados, como escolher a tarefa de mineração adequada e como analisar os resultados obtidos. Ela foi criada com o intuito de abstrair boa parte do conhecimento exigido do usuário para executar o processo envolvido na mineração de dados.

Sua arquitetura é composta por três módulos principais: **Módulo de Apoio à Origem**, **Módulo de Apoio à Preparação** e **Módulo de Apoio à Análise**, conforme observado na Figura 1, retirada de (MENDES, 2009). Em sua interface, é fornecida uma ajuda lateral, que busca auxiliar o usuário na execução das etapas do processo envolvido na mineração de dados.

No **módulo de apoio à origem** são identificadas as fontes de dados que serão utilizadas.

No **módulo de apoio à preparação** são executadas todas as atividades de preparação de dados, que contempla a integração, limpeza, seleção e transformação dos dados.

No **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.

Conforme Vieira *et al*.(2009), com o uso da ferramenta Kira em salas de aula, observou-se que alunos, tanto de graduação quanto graduados, tem 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.

2. Objetivos

Este trabalho objetivou incorporar um algoritmo de classificação na ferramenta de mineração de dados Kira. Esta ferramenta tem como objetivo guiar o usuário, com pouco conhecimento no Processo de KDD e suas etapas, para a execução do processo envolvido na mineração.

Portanto, a incorporação do algoritmo teve por princípio auxiliar o usuário na mineração de dados usando a tarefa de classificação, guiando-o durante todo o processo. Para que isso fosse possível, foram realizados diversos estudos e elaboradas interfaces da forma mais intuitiva possível. Por fim, foi incorporado o módulo de classificação desenvolvido no projeto de pesquisa de um trabalho de mestrado da Universidade Metodista de Piracicaba; alguns ajustes foram feitos em sua interface e em algumas funcionalidades, buscando a melhor maneira de guiar o usuário na execução da mineração de dados com a tarefa de classificação.

3. Desenvolvimento

Para a incorporação do algoritmo de classificação na ferramenta, foi escolhido o algoritmo de Christian Borgelt (BORGELT, 2003).
Na primeira parte do projeto foram realizados diversos experimentos, com conjuntos de dados distintos, utilizando diferentes algoritmos da ferramenta Weka, além do algoritmo de Borgelt (2003). Na segunda parte do projeto buscou-se a familiarização com o código da ferramenta Kira, com a linguagem de programação Java e com o gerenciador de banco de dados Firebird. Foi incorporado, na ferramenta Kira, o algoritmo de Borgelt (2003), com base no trabalho de Coelho (2010); com isso, foram realizados testes adotando diferentes conjuntos de dados e efetuados alguns ajustes. Todo o processo de incorporação da classificação foi realizado com o devido cuidado de desenvolver interfaces de fácil assimilação por parte do usuário.

4. Resultado e Discussão

Como citado anteriormente, neste projeto foi incorporado o módulo de classificação na Kira, com base no trabalho de Coelho (2010). Algumas mudanças foram feitas, consideradas necessárias; elas tiveram como objetivo prover uma melhor forma de ensinar a executar a tarefa de classificação; isso exigiu um conhecimento mais profundo da classificação, obtido durante o desenvolvimento deste projeto de pesquisa.
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 e o objetivo a ser alcançado. Essas funcionalidades são as mesmas da tarefa de associação, 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.
Na etapa de Seleção dos Dados foram desenvolvidos recursos para o usuário selecionar as tabelas, colunas e indicar o atributo classe que será usado.
Depois de selecionar as colunas a serem utilizadas e o atributo classe, foi disponibilizada a opção para o usuário visualizar a árvore de decisão gerada, as regras de classificação extraídas e a matriz de confusão, como forma de avaliação do modelo criado.
No módulo incorporado, foi disponibilizada a leitura das regras geradas, facilitando a interpretação, por parte do usuário.
Na parte 1 da Figura 2 é mostrada a tela referente a exibição das regras de classificação geradas. Um exemplo de modificação realizada na interface desenvolvida por Coelho (2010) é a leitura da regra, exibida na parte 2 da Figura 2, que foi ajustada.
A Figura 3 é referente a uma outra interface pertencente ao módulo de classificação incorporado na ferramenta; a tela mostra a exibição da árvore de decisão gerada a partir dos dados submetidos ao algoritmo de classificação de Borgelt (2003).
Foram realizadas diversas modificações nas interfaces originais do módulo desenvolvido por Coelho (2010). Essas modificações foram realizadas com base em discussões em grupo, visando obter interfaces mais amigáveis aos usuários. Vários protótipos de interfaces foram feitos até se obter aquelas consideradas mais adequadas.

5. Considerações Finais

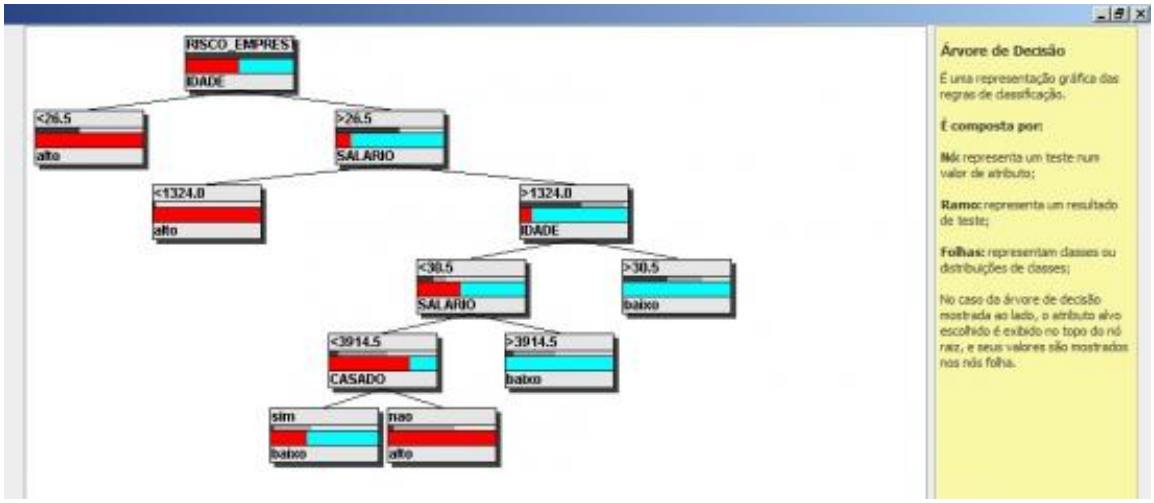
Durante a revisão bibliográfica obteve-se um conhecimento teórico sobre o processo envolvido na mineração de dados; isso foi importante para estabelecer os passos que o usuário deveria seguir para executar satisfatoriamente a etapa da mineração de dados utilizando a tarefa de classificação.
Outro fator relevante na especificação do módulo de classificação incorporado na ferramenta foi decorrente das discussões efetuadas com o grupo de pesquisa, visando identificar a melhor forma de guiar o usuário, em relação à interface desenvolvida.
Na incorporação da classificação 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 diversas 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 módulo de classificação já foi totalmente incorporado na ferramenta e está funcionando corretamente. Como próximas etapas pretende-se realizar testes com usuários, de modo a identificar necessidades de refinamentos nas instruções que orientam a execução do módulo de classificação.

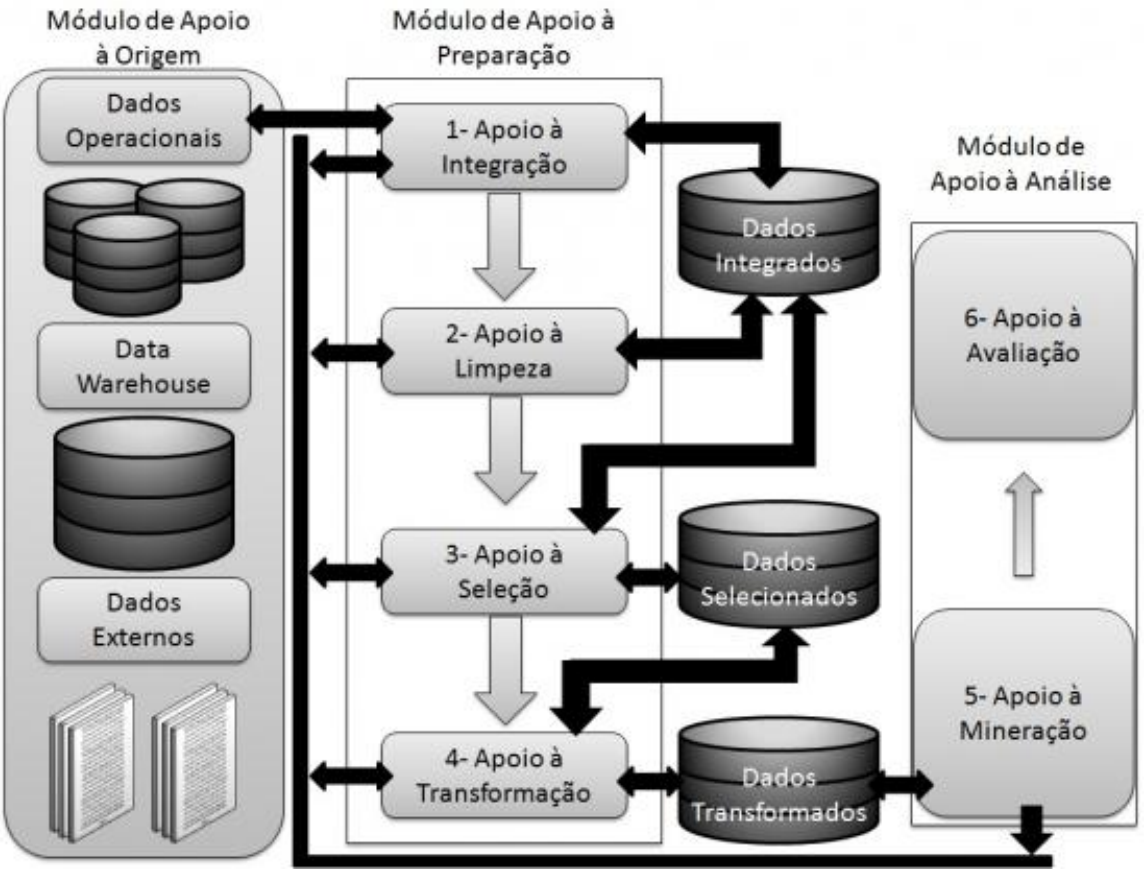
Referências Bibliográficas

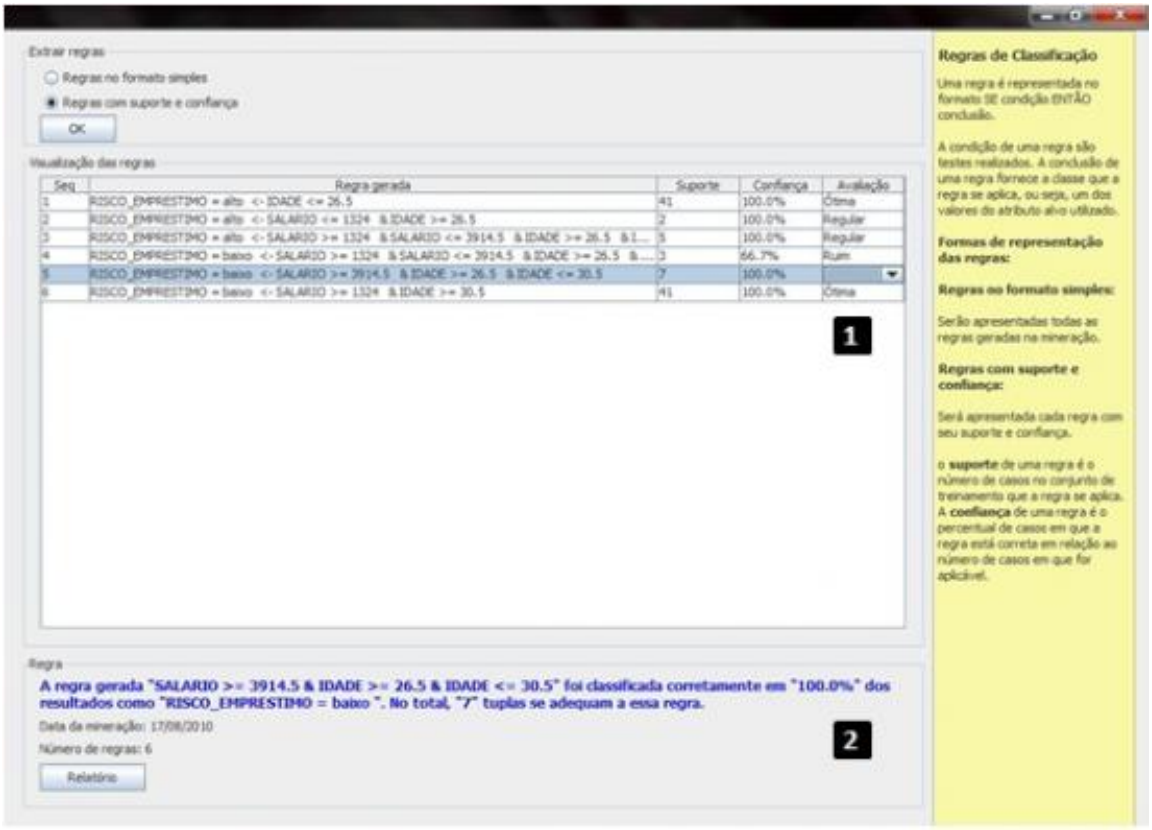
BORGELT, C.. ***Decision and Regression Trees - dti/dtp/dtx/dtr/rsx - induce, prune, and execute decision and regression trees.*** Disponível em: http://www.borgelt.net/doc/dtree/dtree.html. Acesso em: 10 ago. 2010. Copyright © 1996-2003 Christian Borgelt,

2003.
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, E. F.. **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.
MENDES, E. F.; VIEIRA, M. T. P.. **Kira: Uma Ferramenta Instrucional para Apoiar a Aplicação do Processo de Mineração de Dados.** In: SIMPÓSIO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO (SBIE), Florianópolis, p. 1-10, 2009, http://www.br-ie.org/pub/index.php/sbie/article/viewFile/1149/1052.
TAN, P.; STEINBACH, M.; KUMAR, v. - ***Introduction to Data Mining.*** 1a edição. Michigan State University, University of Minnesota, Army High Performance Computing Research Center, USA, 2006.
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. PROCEEDINGS OF THE IADIS INTERNATIONAL CONFERENCE APPLIED COMPUTING 2009, 2009. v. II. p. 12-16.
WITTEN, I.; FRANK, E.. ***Data Mining – Practical Machine Learning Tools and Techniques with Java Implementations.*** 2a edição. San Francisco, USA, 2005.

Anexos




Módulo de Apoio à Origem
Módulo de Apoio à Preparação
Módulo de Apoio à Análise
Dados Operacionais
Data Warehouse
Dados Externos
1- Apoio à Integração
2- Apoio à Limpeza
3- Apoio à Seleção
4- Apoio à Transformação
Dados Integrados
Dados Selecionados
Dados Transformados
6- Apoio à Avaliação
5- Apoio à Mineração


Extrair regras
Regras no formato simples
Regras com suporte e confiança
OK
Visualização das regras
Seq
Regra gerada
Suporte
Confiança
Avaliação
1
RISCO_EMPRESTIMO = alto <- IDADE <= 26.5
41
100.0%
Ótima
2
RISCO_EMPRESTIMO = alto <- SALARIO <= 1324 & IDADE >= 26.5
2
100.0%
Regular
3
RISCO_EMPRESTIMO = alto <- SALARIO >= 1324 & SALARIO <= 3914.5 & IDADE >= 26.5 &1...
3
100.0%
Regular
4
RISCO_EMPRESTIMO = baixo <- SALARIO >= 1324 & SALARIO <= 3914.5 & IDADE >= 26.5 &...
3
66.7%
Ruim
5
RISCO_EMPRESTIMO = baixo <- SALARIO >= 3914.5 & IDADE >= 26.5 & IDADE <= 30.5
7
100.0%
6
RISCO_EMPRESTIMO = baixo <- SALARIO >= 1324 & IDADE >= 30.5
41
100.0%
Ótima
1
Regra
A regra gerada "SALARIO >= 3914.5 & IDADE >= 26.5 & IDADE <= 30.5" foi classificada corretamente em "100.0%" dos resultados como "RISCO_EMPRESTIMO = baixo ". No total, "7" tuplas se adequam a essa regra.
Data da mineração: 17/08/2010
Número de regras: 6
Relatório
2
Regras de Classificação
Uma regra é representada no formato SE condição ENTÃO conclusão.
A condição de uma regra são testes realizados. A conclusão de uma regra fornece a classe que a regra se aplica, ou seja, um dos valores do atributo alvo utilizado.
Formas de representação das regras:
Regras no formato simples:
Serão apresentadas todas as regras geradas na mineração.
Regras com suporte e confiança:
Será apresentada cada regra com seu suporte e confiança.
o suporte de uma regra é o número de casos no conjunto de treinamento que a regra se aplica.
A confiança de uma regra é o percentual de casos em que a regra está correta em relação ao número de casos em que for aplicável.