Estudo de Caso – Regras de Classificação

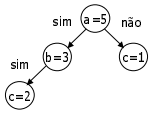
Classificação

Consiste na generalização de exemplos com suas respectivas classes conhecidas em um modelo capaz de reconhecer a classe de um novo exemplo:

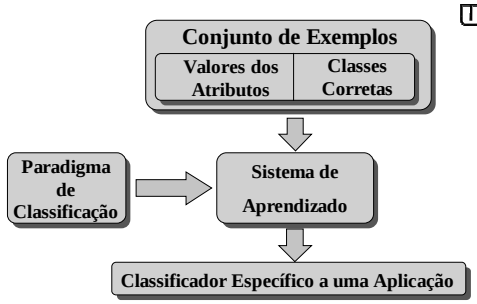
Regras de Produção

a=5 **and** b=3 **->** c=2

Árvores de Decisão



Classificação



Classificação

Estudo de Caso

- Conhecimento do Domínio
- Identificação do Problema
- Descrição do Conjunto de Dados
- Pré-Processamento
- Extração de Padrões
- Pós-Processamento

Conhecimento do Domínio

Os dados disponibilizados para o *Discovery Challenge* foram fornecidos por alguma entidade financeira, contendo informações sobre as contas de seus clientes.

Identificação do Problema

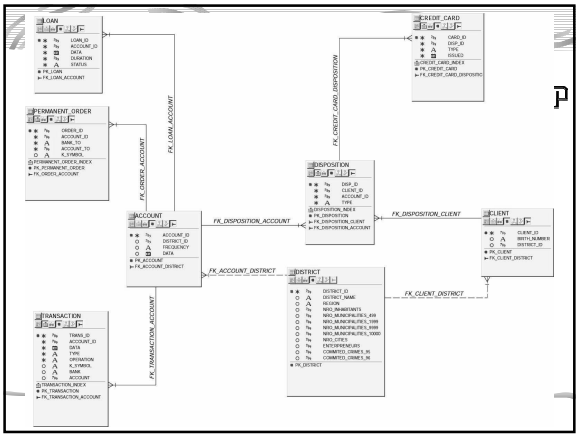
LABIC USP

O objetivo da análise dos dados é verificar os empréstimos oferecidos pelo banco, visando identificar as características dos clientes bons e dos clientes ruins.

Descrição do Conjunto de Dados

LABIC USP

Tabela	Número de Registros	Descrição
account	4500	Contém características estáticas de cada conta
loan	682	Cada registro representa um empréstimo relacionado com uma determinada conta
permanent_order	6471	Os registros apresentam características das ordens de pagamentos
transaction	1056320	Cada registro representa uma transação de uma determinada conta
credit_card	892	Contém os cartões de crédito relacionados com as contas
disposition	5369	Os registros apresentam os direitos de cada cliente para operar sua conta
district	77	Cada registro apresenta informações demográficas das regiões em que os clientes residem
client	5369	Apresenta as características dos clientes



Descrição do Conjunto de Dados

LABIC USP

- Account

item	meaning	remark
account_id	Identificação da conta	
district_id	Localização da agência	
date	Data de criação da conta	Na forma: AAMMDD
frequency	Frequência de emissão de extrato	"POPLATEK MESICNE" (emissão mensal) "POPLATEK TYDNE" (emissão semanal) "POPLATEK PO OBRATU" (emissão depois da transação)

Descrição do Conjunto de Dados

LABIC USP

- Client

item	meaning	remark
client_id	Identificação do cliente	
birth_number	Aniversário e sexo	Homens: na forma AAMMDD Mulheres: na forma AAMM+50DD, Onde AAMMDD é a data de nascimento
district_id	Endereço do cliente	

Descrição do Conjunto de Dados

LABIC USP

- Disposition

item	meaning	remark
disp_id	Identificação do registro	
client_id	Identificação de um cliente	
account_id	Identificação de uma conta	
type	Tipo (proprietário/usuário)	Somente proprietários podem, por exemplo, pedir empréstimo

Descrição do Conjunto de Dados

LABIC USP

- Permanent order

item	meaning	remark
order_id	Identificação do registro	
account_id	Conta para a qual a ordem é emitida	
bank_to	Banco do receptor	Cada banco tem um código de 2 letras
account_to	Conta do receptor	
amount	Quantia debitada	
k_symbol	Característica do pagamento	"POJISTINE" (pagamento de seguro) "SIPO" (pagamento de casa) "LEASING" (leasing) "UVER" (pagamento de empréstimo)

Descrição do Conjunto de Dados

LABIC USP

- Transaction

item	meaning	remark
trans_id	Identificação do registro	
account_id	Conta sobre a qual é feita a transação	
date	Data da transação	Na forma AAMMDD
type	Crédito ou débito	"PRIJEM" (crédito) "VYDAJ" (débito)
operation	Modo da transação	"VYBER KARTOU" (débito com cartão de crédito) "VKLAD" (crédito com dinheiro) "PREVOD Z UCTU" (recebimento de outro banco) "VYBER" (débito em dinheiro) "PREVOD NA UCET" (transferência para outro banco)

Descrição do Conjunto de Dados

LABIC USP

- Transaction (cont.)

amount	Quantia	
balance	Balanço depois da transação	
k_symbol	Característica da transação	"POJISTINE" (pagamento de seguro) "SLUZBY" (pagamento por extrato) "UROK" (crédito) "SANKC. UROK" (aprovação de benefício se o balanço é negativo) "SIPO" (pagamento de casa) "DUCHOD" (aposentadoria) "UVER" (pagamento de empréstimo)
bank	Banco do parceiro	Cada banco tem um código de 2 letras
account	Conta do parceiro	

Descrição do Conjunto de Dados

LABIC USP

- Loan

item	meaning	remark
loan_id	Identificação do registro	
account_id	Identificação da conta	
date	Data em que o empréstimo foi concedido	Na forma: AAMMDD
amount	Quantia	
duration	Duração do empréstimo	
payments	Pagamento mensal	
status	Status do pagamento	'A' (Fim do contrato, sem problemas) 'B' (Fim do contrato, cliente em débito) 'C' (Contrato não finalizado, cliente OK) 'D' (Contrato não finalizado, cliente não está em dia)

Descrição do Conjunto de Dados

LABIC USP

- Credit card

item	meaning	remark
card_id	Identificação do registro	
disp_id	Disposição para uma conta	
type	Tipo do cartão	Valores possíveis: "junior", "classic", "gold"
issued	Data de emissão	Na forma: AAMMDD
A1 = district_id	Código do distrito	
A2	Nome do distrito	
A3	Região	
A4	Número de habitantes	

Descrição do Conjunto de Dados

LABIC USP

- Credit card (cont.)

A5	Número de municípios com nº de habitantes < 499	
A6	Número de municípios com nº de habitantes 500-1999	
A7	Número de municípios com nº de habitantes 2000-9999	
A8	Número de municípios com nº de habitantes > 10000	
A9	Número de cidades	
A10	Proporção de habitantes urbanos	
A11	Salário médio	

Descrição do Conjunto de Dados

- Credit card (cont.)

A12	Taxa de desemprego em 95	
A13	Taxa de desemprego em 96	
A14	Número de empresários por 1000 habitantes	
A15	Número de crimes cometidos em 95	
A16	Número de crimes cometidos em 96	

Pré-Processamento dos Dados

- **Inserção na Base de Dados MySql**

  Os dados foram fornecidos em formato texto. Para facilitar seu pré-processamento, essas tabelas foram criadas em uma base de dados MySql.

Pré-Processamento dos Dados

- **Junção das Tabelas**

  Para fazer a junção dos dados fornecidos, foi utilizada a coluna *status* da tabela *loan*, que classifica os clientes como bons ou ruins.

  Na geração da tabela *Financial_Data_1*, foram utilizados os atributos de todas as tabelas da base e alguns que foram criados durante o pré-processamento.

Pré-Processamento dos Dados

- **Transformação de Valores de Atributos**

  O atributo *account_frequency* apresentava os seguintes valores: *'POPLATEK MESICNE, POPLATEK TYDNE ou POPLATEK PO OBRATU* que foram substituídos por *monthly, weekly ou after transaction*.

  As seguintes colunas também tiveram seus valores alterados: *permanent_order_ksymbol, transaction_type, transaction_operation e transaction_k_symbol*.

Pré-Processamento dos Dados

- **Criação de Novos Atributos**

  Dado o objetivo da análise, foi criada a coluna *loan_predict*. (atributo cujo valor deve ser predito pelas regras extraídas):

  loan_predict = *GOOD* se *loan_status* = *A*,
  loan_predict = *GOOD* se *loan_status* = *C*,
  loan_predict = *BAD* se *loan_status* = *B* e
  loan_predict = *BAD* se *loan_status* = *D*.

Pré-Processamento dos Dados

Foram também criados os seguintes atributos: *client_age*, *client_sex*, *account_months_before_loan*, *credit_card_months_before_loan* e *transaction_months_before_loan*.

Pré-Processamento dos Dados

- **Redução do Número de Registros**

O conjunto de dados *Financial_Data_1* apresentou um volume de dados não suportado pelos algoritmos de Aprendizado de Máquina. Desta maneira, foi criada uma amostra desse conjunto de dados contendo aproximadamente 30% dos dados originais.

Pré-Processamento dos Dados

- **Geração de Arquivos de Dados na Sintaxe Padrão**

No final do pré-processamento dos dados, eles foram convertidos para a sintaxe padrão do projeto Discover.

Para ambos os conjuntos de exemplos criados, foram gerados arquivos de treinamento (2/3 do conjunto de exemplos) e de teste (1/3 dos exemplos).

Extração de Padrões

Para a extração de padrões foi utilizado o algoritmo de Aprendizado de Máquina C4.5:

c4.5 –arquivo_entrada(sem extensão) -u > arquivo_saída.out

A árvore de decisão induzida do conjunto de dados *Financial_Data_1* apresentou uma taxa de erro estimada de 0.5%, tanto no conjunto de treinamento e como no de testes.

Extração de Padrões

Ao verificar a árvore de decisão induzida, observou-se que não foram utilizados os atributos da tabela *transaction* para realizar a classificação. Assim, foi gerado um segundo conjunto de dados (*Financial_Data_2*), sem utilizar a tabela *transaction* para na junção dos dados.

A árvore de decisão induzida a partir desse conjunto de dados apresentou uma taxa de erro estimada de 7.1%.

Pós-Processamento

No pós-processamento das árvores de decisão, foram utilizados os *scripts* para sua conversão em conjuntos de regras na sintaxe padrão de classificação do projeto Discover.

Posteriormente, foram utilizados os *scripts* para cálculo dos valores da matriz de contingência para cada regra a partir dos conjuntos de dados de testes.

Pós-Processamento

- Análise das Regras Utilizando o *RulEE*

O ambiente *RulEE* foi desenvolvido com objetivo de auxiliar a análise de regras. O ambiente possibilita a execução dessa análise através da utilização de suas funcionalidades em *scripts* de pós-processamento, mas a mesma pode ser realizada mais facilmente pela utilização da interface do ambiente.

Pós-Processamento

- Utilização do Ambiente pela Interface

Para exemplificar a utilização do ambiente *RulEE* pela Interface, foi inserido o conjunto de regras extraído a partir do conjunto de dados *Financial_Data_2*.

Foram também realizadas algumas consultas no conjunto de regras referente a esse conjunto de dados.

Pós-Processamento

As consultas realizadas foram as seguintes:

1. Verificação das características dos cliente bons e ruins. Consulta realizada:

SELECT rule_number, rule_antecedent, rule_consequent FROM 7 ORDER BY rule_consequent, rule_number;

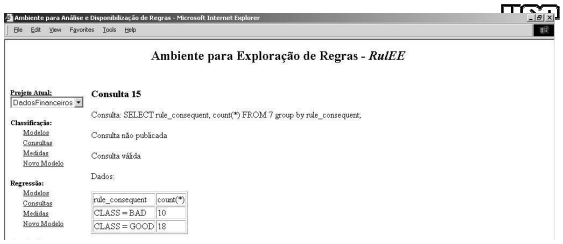
Resultado da consulta 1

Pós-Processamento

2. 18 regras classificam os cliente como GOOD e 10 regras classificam os clientes como BAD. Consulta realizada:

SELECT rule_consequent, count(*) FROM 7 group by rule_consequent;

Pós-Processamento



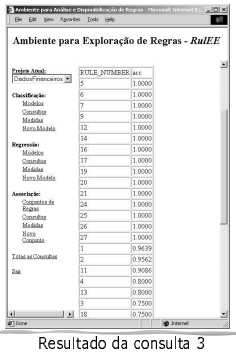
Resultado da consulta 2

Pós-Processamento

3. As regras apresentam uma precisão bastante alta. Consulta realizada:

SELECT RULE_NUMBER, acc FROM 7 ORDER BY acc desc, rule_number asc;

Pós-Processamento



Resultado da consulta 3

Pós-Processamento

4. Pela verificação da precisão relativa, observa-se que as regras que classificam como BAD apresentam, no geral, uma precisão relativa alta. As regras que classificam como GOOD apresentam, no geral, uma precisão relativa baixa. Consulta realizada:

```
SELECT RULE_NUMBER, racc, rule_consequent
FROM 7 ORDER BY racc desc;
```

Pós-Processamento

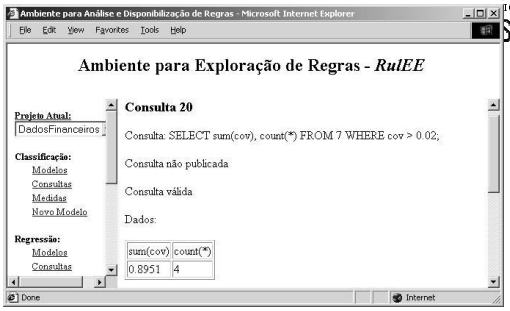
Resultado da consulta 4

Pós-Processamento

5. Poucas regras cobrem uma grande quantidade de exemplos. As 4 regras com maior cobertura (aproximadamente 14% do total) cobrem aproximadamente 90% dos exemplos. Consulta realizada:

```
SELECT sum(cov), count(*) FROM 7 WHERE cov > 0.02;
```

Pós-Processamento

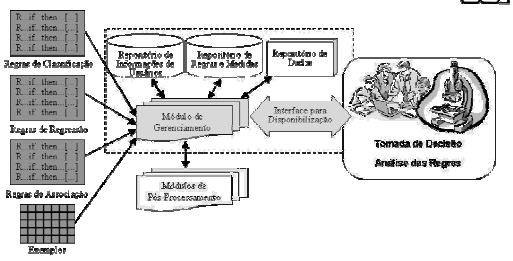


Resultado da consulta 5

Considerações Finais

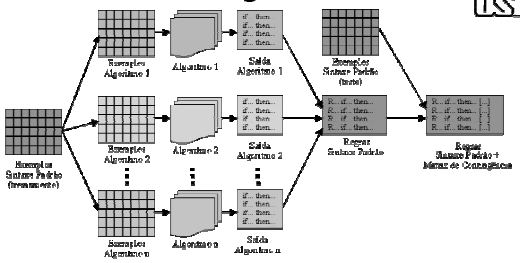
Arquitetura

LABIC USP



Metodologia para Obtenção de Regras

LABIC USP



Referências

LABIC USP

- Paula, M.F., Rezende, S.O.. *Ambiente para Análise e Disponibilização de Regras no Discover*. Resumo apresentado no I Workshop de Teses e Dissertações em Inteligência Artificial do XVI Simpósio Brasileiro em Inteligência Artificial. Porto de Galinhas, 2002
- Paula, M.F., Rezende, S.O.. *Ambiente para Análise e Disponibilização de Regras no Discover*. Resumo apresentado no VII Workshop de Teses e Dissertações em Andamento do VII Simpósio de Teses e Dissertações do Instituto de Ciências Matemáticas e de Computação. São Carlos, 2002

Protótipo

LABIC USP

- http://143.107.232.90/rulee/index.html