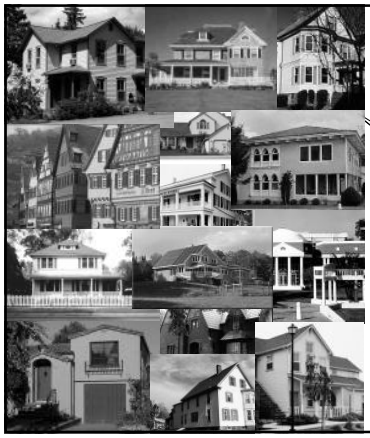
Estudo de Caso

Daniel Gomes Dosualdo
Solange Oliveira Rezende

Índice

- Descrição da Área
- Identificação do Problema
- Descrição do Conjunto de Dados
- Pré-Processamento
- Extração de Padrões
- Pós-Processamento
- Disponibilização do Conhecimento



Descrição da Área

Descrição da Área

- Os experimentos foram realizados tendo como foco a área de imóveis, mais especificamente de casas
- O problema em questão é tentar prever o valor de um imóvel utilizando para isso informações referentes a outros imóveis
- Informações desse tipo são de bastante utilidade para quem trabalha no ramo, principalmente para as imobiliárias

Identificação do Problema

Identificação do Problema

- Baseado em características fornecidas a respeito de imóveis em uma determinada região, o objetivo é tentar predizer o valor de um novo imóvel
- É importante identificar quais as características mais importantes em relação aos imóveis, ou seja, aquelas que possuem peso na tentativa de prever o valor de um novo imóvel

Descrição do Conjunto de Dados Boston Housing

Conjunto de Dados Housing

- Retirado do Repositório de Dados da UCI
- Formado por fatores sócio-econômicos que determinam a compra de imóveis no subúrbio da cidade de Boston, Estados Unidos
- Objetivo é tentar predizer o valor de um imóvel na cidade de Boston

Conjunto de Dados Housing

- Características:
  - Formado por 506 exemplos
  - Possui 14 atributos, todos contínuos
  - Atributo-meta: MedHouseVal
  - Existem valores ausentes

Descrição dos atributos do Conjunto de Dados Housing

Atributo	Descrição
CRIM	taxa de crime por região
ZN	índice de ocupação por área
INDUS	índice de comércio por área
CHAS	atributo simulado Charles River
NOX	concentração de óxidos nítricos
RM	número de quartos por habitação
AGE	proporções de unidades construídas e ocupadas pelo dono antes de 1940
DIS	distância medida para 5 grandes centros de trabalho
RAD	índice de acessibilidade para as rodovias radiais
TAX	taxa sobre o valor do imóvel
PTRATIO	taxa professor-aluno por região
B 1000	proporção de negros por cidade
LSTAT	porcentagem de população pobre
MedHouseVal	valor médio de um imóvel no subúrbio de Boston

Pré Processamento

Pré-Processamento

- Etapa do processo de DM que geralmente consome a maior parte do esforço gasto no processo todo, em torno de 60% do total
- Consiste em preparar os dados para submetê-los a algum algoritmo
- Realiza operações como: adequação de formato, seleção dos dados, transformações, redução da dimensão, dentre outras

Pré-Processamento

- Não foi necessário realizar operações como redução do número de exemplos e redução da dimensão, uma vez que o conjunto de dados Housing é relativamente pequeno
- Em geral, foram efetuadas operações para adequar o conjunto de dados ao formato de entrada dos algoritmos, e algumas operações para retirada de exemplos contendo valores ausentes

Cubist

- Não foi necessária nenhuma operação para adequar os dados retirados da UCI para o formato de entrada do Cubist, uma vez que as sintaxes são compatíveis
- Quanto aos exemplos contendo valores ausentes, o Cubist pode ser executado se for informado os exemplos e atributos em que estão esses valores ausentes

RT

- Para a execução do sistema RT (*Regression Trees*), o arquivo com a extensão *.names* precisou ter sua extensão modificada para *.domain*
- Os exemplos contendo valores ausentes foram retirados para a execução do sistema RT, e portanto, para a execução dos algoritmos RT, CART e RETIS

WEKA

- O WEKA, ao contrário da maioria dos sistemas, recebe como entrada um único arquivo, que contém a declaração dos atributos e os dados propriamente ditos
- Dessa forma, os 2 arquivos originais:
  - receberam o acréscimo das *tags* relation, attribute e data, e
  - tornaram-se um único arquivo, agora com a extensão *.arff*

Extração de Padrões

- Consiste na aplicação de algum algoritmo ao conjunto de exemplos para obter um modelo que represente os padrões extraídos
- Os modelos obtidos a partir de um determinado conjunto de exemplos armazenam o conhecimento adquirido
- Existem muitos tipos de modelos obtidos, como por exemplo, regras e árvores

A Regressão

- No caso de regressão, o modelo obtido descreve uma função de regressão não conhecida
- Esse modelo é então utilizado para predizer o valor de um atributo-meta contínuo de novos exemplos
- Objetivo é encontrar uma relação entre um conjunto de atributos de entrada e um atributo-meta contínuo da seguinte forma:

$$y = f(x_1,x_2,...,x_d)$$

Modelos Baseados em Aprendizado Simbólico

- A maior característica desses modelos é a compreensibilidade dos mesmos
- Variam de acordo com a linguagem escolhida para representar as hipóteses
- Métodos baseados na Lógica Proposicional:
  - Indução de Regras de Regressão
  - Indução *Top-Down* de Árvores de Regressão

Regras de Regressão

- Uma regra de regressão na Forma Normal Conjuntiva é composta de duas partes:
  - a parte condicional das regras, que consiste de uma conjunção de testes realizados nos atributos de entrada, e
  - a parte conclusiva, que contém uma função para predizer o valor do atributo-meta.
- Uma regra na FNC possui a seguinte forma:
  - if <condição> then <y = f(xi)>

Árvores de Regressão

- As árvores são compostas por dois tipos de nós:
  - os nós internos da árvore: cada um desses nós corresponde a um teste feito em um dos atributos de entrada do conjunto de treinamento, e
  - os nós-folha, onde são feitas as predições do atributo-meta.
- Os nós-folha de uma árvore de regressão possuem uma função matemática (que no caso mais simples pode ser a média dos valores que caem em cada nó-folha) para predizer o atributo-meta

Cubist

- O Cubist gera modelos preditivos numéricos baseados em regras
- O modelo construído contém uma ou mais regras, na qual cada regra é uma conjunção de condições associadas com uma expressão linear

Número de regras obtidas pelo Cubist em cada partição

Partição	Número de Regras geradas
1	8
2	9
3	9
4	8
5	8
6	10
7	6
8	9
9	9
10	10

Exemplo de regras geradas pelo Cubist

Rule 1: [60 cases, mean 11.61, range 5 to 20, est err 1.88]

if

CRIM > 5.824

NOX > 0.668

then

MedHouseVal = 18.02 + 3.02 DIS - 0.29 LSTAT - 6 NOX + 0.006 B
- 0.002 TAX + 0.03 RAD - 0.03 CRIM - 0.1 PTRATIO

Rule 2: [22 cases, mean 17.19, range 10.2 to 27.9, est err 4.10]

if

CRIM > 5.824

NOX <= 0.668

LSTAT > 9.71

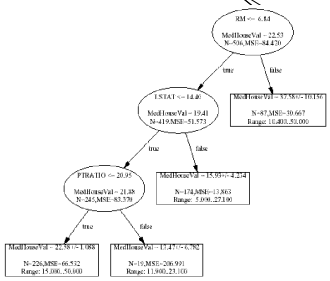
then

MedHouseVal = 32.02 - 0.19 LSTAT - 11 NOX - 0.05 CRIM - 0.19 DIS
+ 0.04 RAD - 0.002 TAX - 0.15 PTRATIO

RT

- O algoritmo RT, implementado por Luis Fernando Torgo, da Universidade de Lisboa, constrói uma árvore de regressão associando valores constantes aos nós-folha da árvore
- A árvore obtida pelo RT sobre o conjunto de dados Housing, após a etapa de poda, possui 145 nós, dos quais 73 são folhas

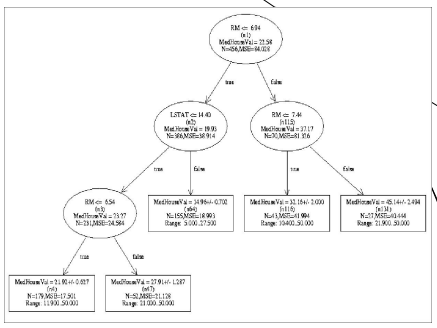
Parte da árvore obtida pelo RT



CART

- O algoritmo CART (*Classification And Regression Trees*) foi desenvolvido por Breiman, Friedman, Olshen e Stone
- Ele permite a construção de árvores de regressão realizando um particionando recursivo binário do conjunto de dados e associando a cada nó-folha da árvore um valor contínuo
- A árvore obtida pelo CART sobre o conjunto de dados Housing possui 9 nós, dos quais 5 são nós-folha

Modelo CART obtido



RETIS

- O algoritmo RETIS (*Regression Tree Induction System*) induzem árvores de regressão que modelam uma relação linear *piecewise* entre atributos discretos ou contínuos e um atributo-meta contínuo
- A árvore obtida pelo algoritmo RETIS sobre o conjunto de dados Housing possui 17 nós, dos quais 9 são nós-folha

M5 Model

- O algoritmo M5 do WEKA, executado com a opção "-m", gera como saída uma árvore com equações lineares associadas aos nós-folha
- A árvore obtida sobre o conjunto de dados Housing com essa opção possui 14 nós-folha

Árvore obtida pelo M5 Model

```
LSTAT <= 9.55 :
|  RM <= 7.13 :
|  |  DIS <= 3.35 : LM1
|  |  DIS >  3.35 : LM2
|  RM >  7.13 :
|  |  RM <= 7.44 :
|  |  |  INDUS <= 5.58 : LM3
|  |  |  INDUS >  5.58 : LM4
|  |  RM >  7.44 :
|  |  |  PTRATIO <= 17.6 : LM5
|  |  |  PTRATIO >  17.6 : LM6
```

```
LSTAT <= 9.55 :
|  RM <= 7.13 :
|  |  DIS <= 3.35 : LM1 (28/77.4%)
|  |  DIS >  3.35 : LM2 (109/31.1%)
|  RM >  7.13 :
|  |  RM <= 7.44 :
|  |  |  INDUS <= 5.58 : LM3 (14/17%)
|  |  |  INDUS >  5.58 : LM4 (5/5.5%)
|  |  RM >  7.44 :
|  |  |  PTRATIO <= 17.6 : LM5 (22/53.3%)
|  |  |  PTRATIO >  17.6 : LM6 (6/52.9%)
```

Algumas equações lineares obtidas pelo M5 Model

- `LM1:  MedHouseVal = 41.3 + 1.22CRIM + 0.016ZN + 0.198CHAS - 6.28NOX + 4.63RM - 5.22DIS + 1.14RAD - 0.0495TAX - 0.171PTRATIO - 0.03B - 0.472LSTAT`
- `LM2:  MedHouseVal = -0.382 + 0.709CRIM + 0.0244ZN + 0.198CHAS - 3NOX + 7.16RM - 0.0327AGE - 0.617DIS + 0.052RAD - 0.015TAX - 0.171PTRATIO - 0.00991B - 0.545LSTAT`
- `LM3:  MedHouseVal = 61.5 - 0.827CRIM + 0.00311ZN + 0.204INDUS + 0.198CHAS - 1.27NOX - 0.584RM - 0.0299AGE - 0.679DIS - 0.163RAD - 0.00857TAX - 0.666PTRATIO + 7.52e-4B - 0.701LSTAT`

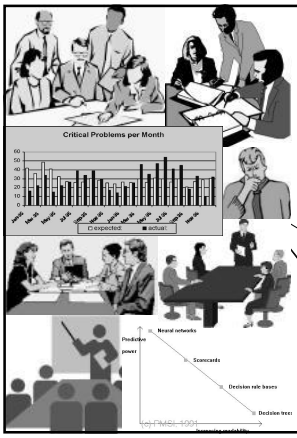
M5 Regression

- Já o algoritmo M5 executado com a opção "-r" gera como saída uma árvore com valores constantes associados aos nós-folha
- A árvore obtida sobre o conjunto de dados Housing nesse caso possui nós, dos quais são nós-folha

M5 Regression

```
LSTAT <= 9.55 :
|  RM <= 7.13 :
|  |  DIS <= 3.35 :
|  |  |  DIS <= 1.94 :
|  |  |  |  DIS <= 1.49 : 50
|  |  |  |  DIS >  1.49 : 29.1
|  |  |  DIS >  1.94 :
|  |  |  |  TAX <= 267 : 32.8
|  |  |  |  TAX >  267 : 23.6
|  |  DIS >  3.35 :
|  |  |  RM <= 6.54 :
|  |  |  |  RM <= 6.06 : 20.6
|  |  |  |  RM >  6.06 : 23.7
|  |  |  RM >  6.54 :
|  |  |  |  LSTAT <= 5.26 : 31.3
|  |  |  |  LSTAT >  5.26 : 26.6
|  RM >  7.13 :
|  |  RM <= 7.44 : 34.6
|  |  RM >  7.44 : 45.3
```

```
LSTAT >  9.55 :
|  LSTAT <= 15 :
|  |  PTRATIO <= 17.9 :
|  |  |  TAX <= 283 : 26.9
|  |  |  TAX >  283 : 21.3
|  |  PTRATIO >  17.9 : 20.2
|  LSTAT >  15 :
|  |  CRIM <= 5.77 :
|  |  |  CRIM <= 0.654 :
|  |  |  |  DIS <= 1.96 : 14.9
|  |  |  |  DIS >  1.96 : 19.7
|  |  |  CRIM >  0.654 : 15.5
|  |  CRIM >  5.77 : 12
```



Pós Processamento

Pós-Processamento

- Os modelos gerados devem ser avaliados para verificar se o conhecimento obtido é novo ou já existente
- Deve-se também verificar, se o conhecimento for novo, se ele é interessante, válido e útil para os usuários do processo

Precisão dos Algoritmos

- Foram utilizadas as medidas MAD e MSE para calcular a precisão dos algoritmos
- A medida MAD (*Mean Absolute Deviation*) consiste na média da diferença (em módulo) entre os valores reais e preditos para um atributo-meta
- A medida MSE (*Mean Squared Error*) consiste na média da diferença ao quadrado entre os valores reais e preditos para um atributo-meta

Precisão dos Algoritmos

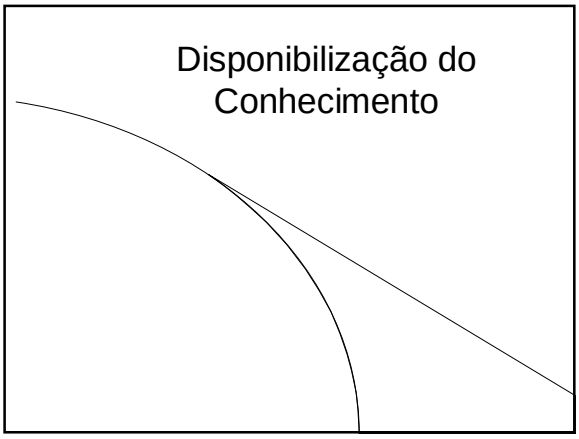
Algoritmo	MAD +/- DP	MSE +/- DP
Cubist	2,770000 +/- 0,155485	12,74135 +/- 3,634049
M5 Model	2,372650 +/- 0,062776	11,85571 +/- 1,045649
M5 Regression	2,824000 +/- 0,056416	16,94432 +/- 1,402621
RETIS	2,866421 +/- 0,111809	17,62673 +/- 2,078216
RT	2,892222 +/- 0,143105	18,72341 +/- 2,424271
CART	3,589523 +/- 0,150907	27,07027 +/- 2,324346

Comparação dos Algoritmos

- Foi realizado um teste para comparação dos regressores, com grau de confiança de 95% assumindo uma distribuição normal
- Constatou-se que:
  - Nenhum algoritmo executado é melhor que todos os demais com um grau de confiança de 95%, baseado na medida MAD

Pós-Processamento

- Nesse caso, foi levado em consideração apenas a precisão dos algoritmos quando executados sobre o conjunto de daddos Housing
- No entanto, existem outros fatores, como a compreensibilidade e interessabilidade dos modelos gerados, que também devem ser levados em consideração


Disponibilização do
Conhecimento