UNIVERSIDADE DE SÃO PAULO

Instituto de Ciências Matemáticas e de Computação

ISSN 0103-2569

Matriz de Contingência para Regras de Regressão

Jaqueline Brigladori Pugliesi

Solange Oliveira Rezende

Nº 210

RELATÓRIOS TÉCNICOS

São Carlos — SP
Agosto/2003

Instituto de Ciências Matemáticas e de Computação

ISSN - 0103-2569

Matriz de Contingência para Regras de Regressão

Jaqueline Brigladori Pugliesi
Solange Oliveira Rezende

RELATÓRIOS TÉCNICOS DO ICMC

São Carlos
Agosto/2003

Matriz de Contingência para Regras de Regressão*

Jaqueline Brigladori Pugliesi
Solange Oliveira Rezende

Universidade de São Paulo
Instituto de Ciências Matemáticas e de Computação
Departamento de Ciências de Computação e Estatística
Laboratório de Inteligência Computacional
Caixa Postal 668, 13560-970 - São Carlos, SP, Brasil
e-mail: {jbpuglie, solange}@icmc.usp.br

Resumo

Diversas medidas disponíveis para avaliação de regras são derivadas da matriz de contingência. Porém, para problemas de regressão o cálculo dessa matriz não é trivial. Neste relatório técnico são apresentadas estratégias para calcular essa matriz e analisar o erro cometido pelas regras de regressão utilizando diferentes medidas. Além disso, neste relatório é feita uma comparação das mesmas com uma medida geralmente utilizada para regressão. Resultados experimentais mostram que o erro baseado na matriz de contingência utilizando uma das estratégias propostas tem comportamento semelhante ao erro calculado sem utilizar a matriz. Com isso, todas as medidas para avaliação de conhecimento baseadas na matriz de contingência podem ser usadas em regras de regressão.

Palavras-Chave: Mineração de Dados, Pós-Processamento, Matriz de Contingência, Regras de Regressão.

Agosto 2003

*Trabalho realizado com auxílio da FAPESP.

Sumário

1 Introdução **1**

2 Regras de Regressão **2**

3 Estratégias para Determinar Matriz de Contingência para Regras de Regressão **3**

3.1 Ajustamento 4

3.2 Pseudo-Classe 5

3.3 Discretização Baseada em Entropia 7

4 Avaliação Experimental **8**

4.1 Experimentos com a Base de Dados Abalone 10

4.2 Experimentos com a Base de Dados Auto-Mpg 11

4.3 Experimentos com a Base de Dados Housing 12

5 Considerações Finais **17**

Referências **20**

Lista de Figuras

1 Desempenho de acordo com o número de pseudo-classes 6
2 Gráfico de quatro medidas para as regras da Base de Dados Abalone. . . . 11
3 Gráfico de quatro medidas para as regras da Base de Dados Auto-Mpg. . . 12
4 Gráfico de quatro medidas para as regras da Base de Dados Housing. . . . 13
5 Boxplot da Base de Dados Abalone. 14
6 Boxplot da Base de Dados Auto-Mpg. 15
7 Boxplot da Base de Dados Housing. 16

Lista de Tabelas

1 Matriz de Contingência para a Regra $B \rightarrow H$. 3
2 Exemplos cobertos pela regra. 5
3 Exemplos não cobertos pela regra. 5
4 Limiares ϵ para a Regra [R0001]. 6
5 Valores para as diferentes medidas de erro para regras da Base de Dados Abalone. 10
6 Valores para as diferentes medidas de erro para regras da Base de Dados Auto-Mpg. 11
7 Valores para as diferentes medidas de erro para regras da Base de Dados Housing. 12

[1]Este documento foi elaborado com o LaTeXe sua bibliografia gerada automaticamente pelo BibTeX.

1 Introdução

O processo de Mineração de Dados inicia-se com a escolha das fontes de dados a serem utilizadas e a definição dos objetivos. Uma parcela desses dados é selecionada, pré-processada e submetida aos métodos e ferramentas adequados com o objetivo de encontrar padrões e/ou modelos que representem o conhecimento obtido. Depois de extraídos, os padrões são pós-processados e o conhecimento adquirido é avaliado quanto à sua qualidade e/ou utilidade para determinar a viabilidade de sua utilização no apoio a algum processo de tomada de decisão.

Quando o conhecimento é obtido de uma Base de Dados, o usuário pode querer saber: se o conhecimento representa o que ele sabe; se não representa, qual a parte do seu conhecimento anterior está ou não correta ou é interessante; ou ainda, de que maneira esse novo conhecimento difere de seu conhecimento anterior (Liu & Hsu 1996; Liu, Hsu, & Chen 1997). Anteriormente, pesquisas assumiam que era de responsabilidade do usuário analisar o conhecimento. Entretanto, quando o número de regras é grande, é muito difícil para este analisá-las manualmente.

A fim de verificar se o conhecimento possui as características desejadas, após adquirido, um pós-processamento deve ser realizado para avaliar o desempenho e a qualidade do mesmo. Essa avaliação procura investigar a precisão e a representação do modelo, a complexidade e a dificuldade de entendimento do conhecimento extraído. Além disso, o pós-processamento também tem como objetivo a filtragem do que foi aprendido, eliminando o conhecimento espúrio, sem valor ou de situações óbvias guiadas pelo senso comum.

A avaliação de conhecimento leva em consideração algumas particularidades dos diferentes tipos de problema. Regras de regressão, apesar de serem semelhantes às regras de classificação, não predizem um valor categórico definido e sim, um valor numérico, envolvendo um grau de incerteza quanto à classe predita (Cheng 1998).

Assim, por ser o atributo meta contínuo, o cálculo da cobertura da regra, ou seja, a verificação da classe estar correta ou não em problemas de regressão se torna mais complicada do que em classificação. Essa verificação é imprescindível para o cálculo dos valores da matriz de contingência das regras. Várias medidas, tanto de precisão quanto de qualidade do conhecimento obtido, são calculadas a partir dos valores da matriz de contingência, como o *framework* descrito em (Lavrač, Flach, & Zupan 1999). Com o cálculo da matriz de contingência para regras de regressão incrementa-se o número de medidas disponíveis para análise das regras.

O foco deste relatório técnico é a utilização de três diferentes estratégias para calcular a matriz de contingência para regras de regressão. As estratégias são: (i) ajustamento, que visa encontrar um limiar entre o valor real e o valor predito, sendo essa estratégia por nós definida; (ii) pseudo-classe, que divide o intervalo de variação da variável meta em pseudo-classes e verifica se o valor predito e o real estão ou não na mesma pseudo-classe; e (iii) discretização baseada em entropia, que também constrói intervalos para a variável meta e verifica se o valor predito e o real se encontram no mesmo intervalo.

Para tanto, é feita uma avaliação experimental comparando os diferentes resultados

da precisão obtidos utilizando a matriz de contingência, calculada por meio dessas três estratégias, com a precisão utilizando a medida de erro MAD (*Mean Absolute Deviation* ou Média da Diferença Absoluta) normalmente usada em avaliação de regras de regressão.

2 Regras de Regressão

No mundo real, problemas de classificação são mais encontrados do que problemas de regressão, o que explica a maior atenção dada a classificação do que a regressão. Porém, muitos problemas importantes são de regressão. Além do fato de que problemas de regressão são importantes por si próprios, outra razão da necessidade em focar em regressão é que métodos de regressão podem ser usados para resolver problemas de classificação. Por exemplo, redes neurais são muitas vezes aplicadas a problemas de classificação (Weiss & Indurkhya 1995).

Em regressão, a cada exemplo é atribuído um valor numérico ao invés de um conjunto finito de rótulos simbólicos. Ou seja, dada uma amostra de um conjunto de variáveis de entrada $x_1, ..., x_n$ junto com a variável meta y, tenta-se aproximar a função, mostrada na Equação 1.

$$y = f(x_1, ..., x_n) \tag{1}$$

Um problema de regressão é, às vezes, descrito como um problema de sinal e ruído. Assim, o modelo é estendido para incluir um componente estocástico ou limiar ϵ (Equação 2). Nesse caso, a função real não pode produzir uma distância de erro zero. Ao contrário da classificação, na qual os rótulos são assumidos corretos, para regressão os valores preditos de y podem ser explicados por alguns fatores incluindo um componente de ruído aleatório, ϵ, no sinal, y.

$$y = f(x_1, ..., x_n) \pm \epsilon \tag{2}$$

Como este trabalho está relacionado com a avaliação do conhecimento de regras de regressão, considere uma regra, que é geralmente representada da seguinte forma:

$$if \underbrace{< condicao >}_{B-Body} then \underbrace{< classe = C_i >}_{H-Head}$$

sendo $< condicao >$ uma disjunção de conjunções para os atributos e C_i um valor numérico ou uma equação.

Para as medidas de avaliação aqui relatadas, será adotada a representação mais genérica ($B \rightarrow H$) a fim de que possa abranger uma maior variedade de regras.

A matriz de contingência pode ser construída, para qualquer regra, a partir do conjunto de dados. Na Tabela 1, B denota o conjunto de exemplos para o qual o corpo ($Body$) da regra é verdadeiro, e $\overline{B}$ denota o seu complemento (o conjunto de exemplos para o qual o

corpo da regra é falso); e de forma similar H e $\overline{H}$. Assim, HB denota $H \cap B$, $\overline{H}B$ denota $\overline{H} \cap B$, e assim por diante. Utilizou-se o $n(X)$ para denotar a cardinalidade do conjunto X, por exemplo, $n(\overline{H}B)$ é o número de exemplos para o qual H é falso e B é verdadeiro (ou seja, o número de exemplos erroneamente cobertos pela regra). A freqüência relativa $\frac{n(X)}{N}$ associada a X é denotada por $p(X)$.

Tabela 1: Matriz de Contingência para a Regra $B \rightarrow H$.

	B	$\overline{B}$	
H	$n(HB)$	$n(H\overline{B})$	$n(H)$
$\overline{H}$	$n(\overline{H}B)$	$n(\overline{H}\overline{B})$	$n(\overline{H})$
	$n(B)$	$n(\overline{B})$	N

Dessa maneira, em problemas de regressão, o cálculo do valor do número de exemplos falsos e verdadeiros quanto ao atributo meta (ou seja, a cobertura da regra) se torna mais complicado por ser contínuo, diferentemente da classificação, na qual o atributo meta é discreto e a verificação da classe ser correta ou não é direta. Ou seja, o cálculo dos valores relacionados à H e $\overline{H}$ não são triviais, sendo um problema que se encontra ainda em aberto na comunidade da área.

Levando-se em consideração também que várias medidas de avaliação de conhecimento utilizam em suas fórmulas algum dos valores pertencentes à matriz de contingência, ressalta-se que o cálculo dessa matriz é muito importante (Mitchell 1997; Lavrač, Flach, & Zupan 1999).

Outras medidas, utilizadas em problemas de regressão, não levam em consideração os valores da matriz de contingência, como algumas medidas para calcular o erro: MAD (*Mean Absolute Deviation* ou Média da Diferença Absoluta), MSE (*Mean Squared Error* ou Média dos Erros ao Quadrado) e RMSE (*Relative Mean Squared Error* ou MSE Relativa) (Torgo 1995; Merz 1998; Torgo 1999).

Neste trabalho foram utilizadas três estratégias para o cálculo dos valores da matriz de contingência para problemas de regressão. Com isso, as medidas baseadas na matriz de contingência também podem ser aplicadas para avaliar regras de regressão.

3 Estratégias para Determinar Matriz de Contingência para Regras de Regressão

Visando o cálculo dos valores da matriz de contingência foram propostas a utilização de três estratégias. Uma das estratégias, que visa um ajustamento entre o valor real e o predito, foi por nós proposta e implementada. Já as outras duas efetuam uma discretização, gerando intervalos de valores para o atributo meta, de forma a verificar se o resultado encontrado está correto ou não. Dessa maneira, foram utilizados dois critérios para encontrar as classes: pseudo-classe e discretização baseada em entropia. Essas três estratégias serão detalhadas a seguir.

3.1 Ajustamento

Como o algoritmo de regressão retorna como classe ou cabeça (*Head*) da regra um valor numérico real ou uma equação matemática baseada em seus atributos, que gera como resultado final para cada exemplo um valor real, utilizou-se um limiar para a construção da matriz de contingência. Esse limiar (ϵ) é apresentado na Equação 2. Com a utilização do limiar ϵ é possível calcular o número de exemplos para o qual a cabeça da regra é verdadeira ou falsa, ou seja, se o valor referente a classe do exemplo pertence ao intervalo definido por esse limiar.

O ponto crítico nesta estratégia é a definição do valor do limiar. O limiar utilizado na estratégia, por nós definida e implementada, utiliza o cálculo do desvio padrão do erro da regra encontrado no conjunto de dados.

Com o objetivo de analisar como diferentes limiares utilizando desvio padrão influenciam nos valores da matriz de contingência três variações foram consideradas: somente o valor do desvio padrão, o desvio padrão mais a média do erro e o desvio padrão mais a mediana do erro. Com isso é possível analisar o comportamento de diferentes tamanhos de limiar, dado que utilizando somente o desvio padrão tem-se um limiar mais restrito.

Para cada regra são calculados o desvio padrão, a média e a mediana, que leva em consideração tanto os exemplos cobertos por essa regra, quanto os não cobertos. Ou seja, para cada regra tem-se dois desvios padrão, duas médias e duas medianas, que são utilizadas de acordo com a cobertura do exemplo.

Como exemplo, considere a seguinte regra, na sintaxe padrão (Pugliesi, Dosualdo, & Rezende 2003), na qual `rings` é o atributo a ser predito:

```
[R0001]  IF shell_weight <= 0.25
         THEN rings = 5 +21.4 shell_weight +1.7 diameter
```

Suponha que o conjunto de exemplos possui 12 exemplos coberto por essa regra e 12 exemplos não cobertos por essa regra, como são apresentados nas Tabelas 2 e 3, respectivamente.

Para cada um dos exemplos, foi calculado o valor predito pela regra, bem como o erro encontrado, ou seja, a diferença absoluta entre o valor predito e o real. Por fim, foram calculadas as medidas relacionadas a regra, sendo o desvio padrão igual a 1,86; a média igual a 1,36 e a mediana igual a 0,75.

Além disso, para cada um dos exemplos não cobertos pela regra, também foi calculado o valor predito, bem como o erro encontrado. E foram calculadas as medidas relacionadas a essa regra, sendo o desvio padrão igual a 2,36; a média igual a 2,89 e a mediana igual a 2,60.

Como esperado, os valores relacionados aos exemplos não cobertos pela regra apresentaram maiores valores de erro.

Tabela 2: Exemplos cobertos pela regra.

Atributos									**Valor Predito**	**Erro**
sex	length	diameter	height	whole weight	shucked weight	viscera weight	shell weight	rings		
M	0,46	0,37	0,10	0,51	0,22	0,10	0,15	15	8,83	6,17
M	0,35	0,27	0,09	0,23	0,10	0,05	0,07	7	6,95	0,05
F	0,53	0,42	0,14	0,68	0,26	0,14	0,21	9	10,20	1,20
M	0,44	0,37	0,13	0,52	0,22	0,11	0,16	10	8,94	1,06
I	0,33	0,26	0,08	0,21	0,09	0,04	0,06	7	6,61	0,39
I	0,43	0,30	0,10	0,35	0,14	0,08	0,12	8	8,08	0,08
M	0,48	0,37	0,13	0,51	0,22	0,11	0,17	9	9,16	0,16
F	0,53	0,38	0,14	0,61	0,19	0,15	0,21	14	10,10	3,90
M	0,43	0,35	0,11	0,41	0,17	0,08	0,14	10	8,48	1,52
M	0,49	0,38	0,14	0,54	0,22	0,10	0,19	11	9,71	1,29
F	0,54	0,41	0,15	0,68	0,27	0,17	0,21	10	10,10	0,10
F	0,47	0,36	0,10	0,48	0,17	0,08	0,19	10	9,56	0,44

Tabela 3: Exemplos não cobertos pela regra.

Atributos									**Valor Predito**	**Erro**
sex	length	diameter	height	whole weight	shucked weight	viscera weight	shell weight	rings		
F	0,53	0,42	0,15	0,78	0,24	0,14	0,33	20	12,80	7,20
F	0,55	0,43	0,13	0,77	0,29	0,15	0,26	16	11,30	4,70
F	0,55	0,44	0,15	0,89	0,31	0,15	0,32	19	12,60	6,40
F	0,57	0,44	0,16	0,94	0,43	0,21	0,27	12	11,50	0,50
F	0,62	0,48	0,17	1,16	0,51	0,30	0,31	10	12,30	2,30
F	0,56	0,44	0,14	0,93	0,38	0,19	0,30	11	12,20	1,20
F	0,58	0,45	0,19	1,00	0,39	0,27	0,29	11	11,90	0,90
M	0,59	0,45	0,14	0,93	0,36	0,23	0,28	12	11,70	0,30
M	0,61	0,48	0,18	0,94	0,39	0,22	0,30	15	12,10	2,90
M	0,58	0,47	0,17	1,00	0,39	0,24	0,33	10	12,90	2,90
F	0,68	0,56	0,17	1,64	0,61	0,28	0,46	15	15,80	0,80
M	0,67	0,53	0,17	1,34	0,55	0,36	0,35	18	13,40	4,60

Considerando os cálculos apresentados, os limiares ϵ para essa regra podem ser visualizados na Tabela 4. Assim, o exemplo é considerado corretamente predito pela regra caso o valor predito esteja entre o valor real $\pm$ o limiar utilizado.

3.2 Pseudo-Classe

Um conjunto de "pseudo-classes" é gerado utilizando o algoritmo P-Class em um processo de discretização descrito em (Weiss & Indurkhya 1995). O objetivo desse algoritmo é associar os valores do atributo-meta y às pseudo-classes de modo a minimizar a distância entre cada y_i e a média de sua pseudo-classe.

O P-Class descreve como associar um conjunto de valores $\{y_i\}$ a k classes da seguinte maneira:

Tabela 4: Limiares ϵ para a Regra [R0001].

	Desvio Padrão	Desvio Padrão e Média	Desvio Padrão e Mediana
B (shell_weight <= 0.25)	1,86	3,22	2,61
$\overline{B}$	2,36	5,25	4,96

a. ordena os valores de $\{y\}$;

b. associa aproximadamente um número igual de valores de $\{y_i\}$ vizinhos para cada pseudo-classe;

c. move um valor y_i para uma pseudo-classe vizinha quando isso reduz a distância de y_i para a média de sua pseudo-classe;

d. concatena pseudo-classes com médias iguais.

Fornecido o valor de k, o algoritmo P-Class associa de maneira relativamente rápida os y valores de modo que as distâncias totais sejam minimizadas. No entanto, uma questão-chave é como determinar k, ou seja, o número de pseudo-classes a serem geradas. Em (Weiss & Indurkhya 1995) são realizados alguns experimentos para determinar o melhor número de pseudo-classes. Na Figura 1 é apresentado um gráfico do erro relativo *versus* o número de pseudo-classes. Como pode ser visualizado nesse gráfico, à medida que o número de pseudo-classes aumenta, os resultados melhoram, até alcançar o valor 6, sendo que logo em seguida os resultados começam a piorar.

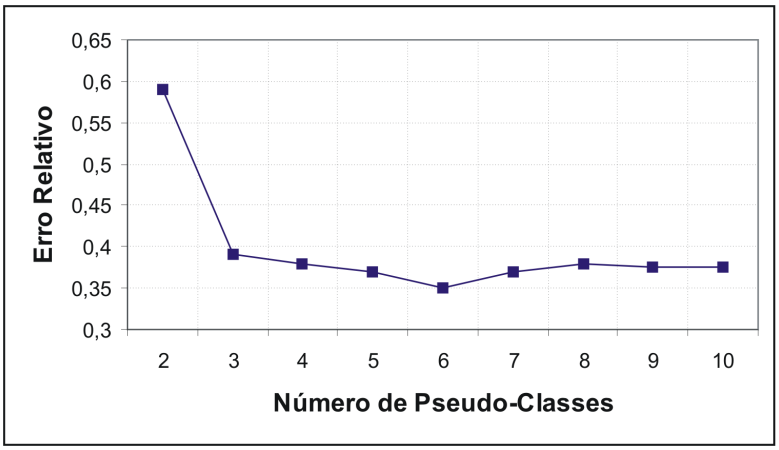


Figura 1: Desempenho de acordo com o número de pseudo-classes (Weiss & Indurkhya 1995)

Esse algoritmo P-Class foi desenvolvido com o intuito de transformar um problema de regressão em um problema de classificação. Para isso, os valores numéricos do atributo-meta passam por um processo de discretização, em que são geradas as pseudo-classes. Realizada a discretização dos valores (geradas as pseudo-classes), um algoritmo de indução de regras de decisão como o C4.5 (Quinlan 1993) ou o CN2 (Clark & Niblett 1989), pode

ser utilizado para gerar as regras. Com as regras induzidas, no processo proposto em (Weiss & Indurkhya 1995) são aplicados então procedimentos de poda e otimização para produzir um conjunto otimizado de regras.

Na estratégia de pseudo-classes, a especificação das classes não utiliza informação nenhuma além da ordem de y (mostrado na Equação 1). Nenhuma suposição sobre a verdadeira natureza da função base é feita. O objetivo, com essa estratégia, é fazer com que a distância entre cada y e a média de sua pseudo-classe seja minimizada. Essa abordagem é relatada em (Weiss & Indurkhya 1995), que a utiliza para transformar o problema de regressão em classificação. Porém, neste artigo utiliza-se essa abordagem com o propósito de calcular os valores da matriz de contingência, verificando se o valor predito e o real pertencem ou não à mesma pseudo-classe, sendo a implementação da mesma realizada por (Dosualdo 2003).

Para a realização dos experimentos aqui apresentados utilizou-se 3 e 6 pseudo-classes. O número 6 foi escolhido, pois (Weiss & Indurkhya 1995) mostra que com esse número de classes foram encontrados os melhores resultados nos testes empíricos realizados. Já o número 3 foi escolhido devido ao fato de estarmos utilizando, inicialmente, conjuntos de dados não muito grandes para a realização dos experimentos.

3.3 Discretização Baseada em Entropia

A estratégia de discretização baseada em entropia também utiliza apenas informações da ordem de y. A discretização considera apenas o conjunto de dados original para construir os intervalos.

Essa estratégia requer que seja selecionado um atributo discreto. Os limites de cada intervalo são escolhidos de modo que a distribuição de atributos com faixas diferentes sejam a mais diferente possível. Essa abordagem continua criando limites que separem as faixas até que nenhum intervalo adicional seja considerado significante.

A ferramenta utilizada para verificar essa estratégia foi a versão 2.6 do MineSet^{TM}da Silicon Graphics (Rathjens 1996). Essa ferramenta incorpora técnicas de Mineração de Dados que auxiliam o analista/especialista na exploração dos dados, e possibilita também discretização baseada em peso e escala uniforme. Para a discretização realizada com essa ferramenta foram utilizados os mesmos números de intervalos da estratégia anterior, ou seja, 3 e 6 intervalos. Além disso, também foram feitos experimentos com o MineSet^{TM}gerando automaticamente a quantidade e os limites dos intervalos.

Na próxima seção serão mostrados os testes realizados para verificar a eficiência e eficácia dessas estratégias, bem como uma comparação entre elas e a medida MAD que não leva em consideração os valores da matriz de contingência.

4 Avaliação Experimental

Para avaliar os resultados das três estratégias propostas foram utilizadas 2 medidas. Uma dessas medidas é a precisão, que é uma medida derivada da matriz de contingência (Equação 3). Essa medida é utilizada para validar os resultados das três estratégias usadas para calcular a matriz de contingência.

$$Precisão = p(H|B) = \frac{p(HB)}{p(B)} = \frac{n(HB)}{n(B)} \tag{3}$$

Já a outra medida, que não considera a matriz de contingência, é a MAD, que quantifica o erro do modelo pela média dos desvios absolutos de suas predições, isto é, a MAD consiste na média da diferença (em módulo) entre os valores reais e preditos para um atributo-meta, como é mostrado na Equação 4.

$$\text{MAD}(R) = \frac{1}{N}\sum_{i=1}^{N} |y_i - y'_i| \tag{4}$$

Nessa equação, y é o valor real do atributo-meta do exemplo, N é o número de exemplos, e y' é o valor predito pelo algoritmo.

Para validar as três estratégias utilizando experimentação empírica, uma questão principal precisa ser respondida:

> Quais erros baseados na matriz de contingência, utilizando as diferentes estratégias, possuem valores semelhantes aos encontrados utilizando a medida MAD?

A tarefa de qualquer algoritmo de regressão é tentar estimar o valor da variável meta de um novo exemplo em função das demais variáveis. O algoritmo de regressão utilizado nestes experimentos iniciais foi o Cubist (Rulequest-Research 2000). O Cubist é uma ferramenta utilizada para a geração de modelos preditivos numéricos baseados em regras a partir de um conjunto de dados fornecido. Esses modelos são utilizados na predição de valores numéricos, ou seja, para problemas de regressão. Cada caso fornecido ao Cubist possui uma variável meta e as variáveis preditoras que são utilizadas para predizer o valor dessa variável meta. O Cubist gera regras não ordenadas. Essas regras estão dispostas em ordem crescente do valor médio dos exemplos preditos. Ele constrói um modelo contendo uma ou mais regras, na qual cada regra é uma conjunção de condições associadas com uma expressão linear. Portanto, os modelos gerados pelo Cubist são modelos lineares *piecewise*. Porém, o Cubist pode também construir modelos múltiplos e pode combinar modelos baseados em regras com modelos baseados em exemplos.

Para a realização dos experimentos foi utilizada o algoritmo de regressão Cubist para encontrar os padrões. Além disso, se utilizou as Bases de Dados Abalone, Auto-Mpg e Housing. Por terem sido retiradas do Repositório da UCI, não foi necessário realizar limpeza nos dados antes de iniciar o processo de extração de padrões.

Os experimentos foram realizados utilizando *10-fold cross-validation* e os resultados de todas as regras geradas foram coletados. Para avaliar as regras foi calculada a precisão a partir da matriz de contingência utilizando as três estratégias propostas, além das variações dentro de cada estratégia, ou seja, a estratégia ajustamento (somente o valor do desvio padrão, o desvio padrão mais a média do erro e o desvio padrão mais a mediana do erro); a pseudo-classe (3 e 6 pseudo-classes) e a discretização baseada em entropia (3 e 6 intervalos e automático). Para todas as regras também foi calculada a MAD como parâmetro chave de comparação.

Para que a comparação entre as medidas pudesse ser realizada, normalizou-se a medida MAD para que seus valores ficassem somente entre 0 e 1, como todas as medidas baseadas na matriz de contingência. Houve a necessidade também de padronizar as medidas de maneira que todas elas retratassem o erro cometido pela regra, uma vez que algumas medidas fornecem como resultado o valor da precisão e não do erro. Vale ressaltar que todas as medidas baseadas na matriz de contingência são comparadas com a MAD, para que se possa verificar quais medidas possuem resultados semelhantes a MAD, de modo a verificar qual das estratégias gerou bons resultados para a matriz.

Para uma avaliação comparativa confiável, utilizou-se em todos os experimentos o Teste t-Student com intervalo de confiança de 95% para a análise estatística dos resultados. O Teste t-Student é um teste estatístico cujo objetivo é testar a igualdade entre duas médias. O teste supõe independência e normalidade das observações. As variâncias dos dois grupos podem ser iguais ou diferentes, havendo alternativas de teste para as duas situações.

As análises estatísticas foram feitas a fim de observar a não existência de uma diferença significativa das medidas encontradas para as regras de regressão. Quando um resultado é estatisticamente significante quer dizer que as diferenças encontradas são grandes o suficiente para não serem atribuídas ao acaso.

Para essa verificação foram realizados testes de hipóteses, através do qual expressa-se determinado parâmetro da população e procura-se evidência para rejeitar ou não a hipótese nula (a da não diferença entre duas variáveis). Para tanto, foram utilizadas as seguintes hipóteses:

H_0: Não existe diferença significativa entre as medidas de erro das regras.

H_1: Existe diferença significativa entre as medidas de erro das regras.

A hipótese nula (H_0) é colocada à prova no teste de hipótese. Em geral indica uma igualdade a ser contestada. A hipótese alternativa é a hipótese que será considerada como aceitável, caso a hipótese nula seja rejeitada (Allen 1990; Magalhães & de Lima 2002).

Para a execução dos testes t-Student e geração de gráficos para melhor visualização dos resultados, utilizou-se o software estatístico MINITAB^{TM}[1]. Foram gerados gráficos de todas as medidas *versus* a medida MAD, de modo a verificar quais não possuem diferença significativa com a MAD.

Os testes foram realizados comparando a MAD com as três estratégias, considerando

[1] http://www.minitab.com

também todas as variações em cada estratégia.

4.1 Experimentos com a Base de Dados Abalone

Uma das bases utilizadas nos experimentos foi a Base de Dados Abalone que é utilizada para obter um modelo para predizer a idade de um abalone (tipo de molusco) a partir de medidas físicas. A idade do abalone é determinada cortando a ostra através do cone, colorindo-a e contando o número de anéis por meio de um microscópio — uma tarefa tediosa e demorada. Outras medidas, as quais são fáceis de serem obtidas, são utilizadas para predizer o número de anéis que determinam a idade. Essa Base de Dados encontra-se no repositório da UCI (Blake & Merz 1998), e possui 4177 exemplos e 8 atributos, sendo 1 nominal e 7 contínuos.

Na Tabela 5 são mostrados alguns dos resultados obtidos para as regras geradas pela ferramenta Cubist. Esses resultados estão relacionados à primeira partição da Base de Dados Abalone. Como já esperado, para cada regra gerada os erros calculados considerando as variações em cada estratégia proposta são diferentes. Cada linha da tabela representa os resultados obtidos para uma regra de regressão.

Tabela 5: Valores para as diferentes medidas de erro para regras da Base de Dados Abalone.

		Ajustamento			Pseudo-Classe		Discretização Baseada em Entropia		
Número Regra	MAD	Desvio Padrão	DP e Média	DP e Mediana	3 pseudo-classes	6 pseudo-classes	3 intervalos Mineset	6 intervalos Mineset	Mineset automático
R0001	0,00	0,34	0,10	0,13	0,22	0,47	0,32	0,56	0,35
R0002	0,34	0,45	0,14	0,18	0,38	0,66	0,42	0,64	0,44
R0003	0,56	0,31	0,12	0,16	0,25	0,55	0,08	0,26	0,44
R0004	1,00	0,44	0,15	0,18	0,34	0,61	0,13	0,30	0,58
R0005	0,81	0,45	0,13	0,19	0,24	0,55	0,05	0,19	0,52

Na Figura 2 é mostrado um gráfico de quatro medidas: a MAD e uma para cada estratégia (desvio padrão, 6 pseudo-classes e 6 intervalos do MineSet^{TM}). De cada estratégia foi escolhida a variação que apresentou o melhor resultado. Cada ponto no gráfico é o valor da medida em questão para uma dada regra. No gráfico são apresentadas algumas das 111 regras de regressão obtidas, sendo as 5 primeiras regras as mesmas apresentadas na Tabela 5.

Porém, pode-se perceber que somente a partir da análise visual do gráfico não se consegue chegar a uma conclusão confiável sobre os resultados, devido, em parte, à grande quantidade de informação nele representada.

Com isso, na Figura 5 são apresentados todos os gráficos *boxplot* da diferença entre a medida MAD e as estratégias para a Bases de Dados Abalone. Pode-se ver que a hipótese H_0 não foi rejeitada somente quando se utiliza a estratégia de desvio padrão, pois H_0 está dentro do intervalo de confiança.

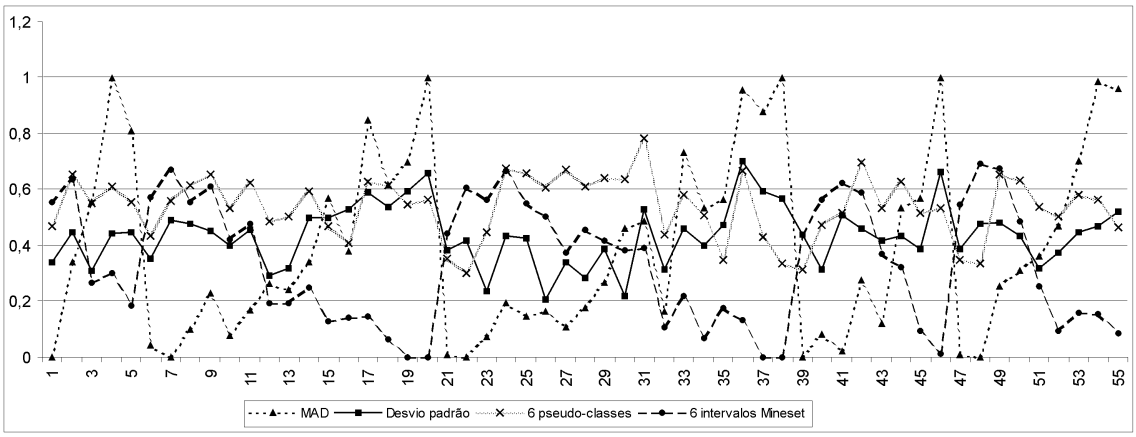


Figura 2: Gráfico de quatro medidas para as regras da Base de Dados Abalone.

4.2 Experimentos com a Base de Dados Auto-Mpg

A Base de Dados Auto-Mpg está relacionada com o consumo de combustível na cidade em mpg (milhas por galão), baseado, entre outros, nos atributos cilindros, poder do motor, peso, aceleração e ano de fabricação. Esta Base de Dados também foi obtida no repositório da UCI, e possui 398 exemplos e 8 atributos, sendo 3 nominais e 5 contínuos. Originalmente, a base possuía 9 atributos, porém um dos atributos, o modelo do carro, é único para todos os exemplos e foi removido.

A Tabela 6 e o gráfico na Figura 3 apresentam os resultados, como os já explicados anteriormente, para a Base de Dados Auto-Mpg. Nesse caso, na Tabela 6 também são apresentadas somente as regras obtidas a partir da primeira partição da Base de Dados Auto-Mpg. Além disso, os resultados das primeiras das 93 regras são apresentados na Figura 3 para as mesmas quatro medidas: MAD, desvio padrão, 6 pseudo-classes e 6 intervalos do Mineset.

Tabela 6: Valores para as diferentes medidas de erro para regras da Base de Dados Auto-Mpg.

		Ajustamento			Pseudo-Classe		Discretização Baseada em Entropia		
Número Regra	MAD	Desvio Padrão	DP e Média	DP e Mediana	3 pseudo-classes	6 pseudo-classes	3 intervalos Mineset	6 intervalos Mineset	Mineset automático
R0001	0,17	0,51	0,14	0,16	0,06	0,35	0,00	0,09	0,32
R0002	0,00	0,58	0,25	0,25	0,08	0,25	0,00	0,17	0,25
R0003	0,18	0,19	0,00	0,05	0,38	0,43	0,00	0,29	0,33
R0004	0,32	0,43	0,14	0,18	0,14	0,39	0,04	0,14	0,43
R0005	0,31	0,50	0,13	0,13	0,06	0,41	0,22	0,22	0,50
R0006	0,56	0,50	0,14	0,14	0,32	0,32	0,27	0,27	0,50
R0007	0,10	0,59	0,13	0,16	0,19	0,22	0,03	0,03	0,31
R0008	0,08	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,75
R0009	0,53	0,63	0,16	0,18	0,03	0,16	0,24	0,26	0,63
R0010	1,00	0,60	0,13	0,13	0,00	0,00	0,40	0,67	0,87

Já na Figura 6 são apresentados todos os gráficos *boxplot* da diferença entre a medida MAD e todas as estratégias para Bases de Dados Auto-Mpg. Nesses casos, observa-se que a hipótese H_0 não foi rejeitada, ou seja, não existe diferença significativa entre as medidas para as estratégias que utiliza o somente o desvio padrão e o mineset automático. Isso

pode ser visto, uma vez que H_0 está contido no intervalo de confiança.

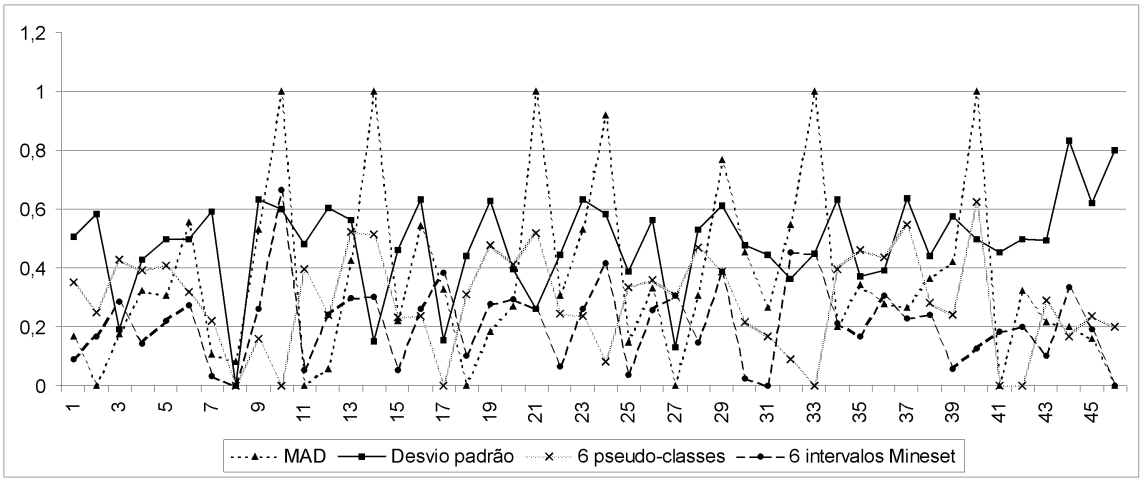


Figura 3: Gráfico de quatro medidas para as regras da Base de Dados Auto-Mpg.

4.3 Experimentos com a Base de Dados Housing

O valor de moradias em subúrbios de Boston pode ser predito com a Base de Dados Housing. Cada exemplo dessa base descreve uma área com características como taxa de crime, concentração de óxido nítrico, acesso a rodovias, número médio de quartos por moradia, etc. A Base de Dados Housing está contida no repositório da UCI, e possui 506 exemplos, 12 atributos contínuos, 1 nominal e a classe numérica, representando o valor das moradias.

A Tabela 7 e o gráfico na Figura 4 apresentam os resultados para a Base de Dados Housing. Nesse caso, na Tabela 7 também são apresentadas somente as regras obtidas a partir da primeira partição da Base de Dados Housing.

Além disso, os resultados das primeiras das 45 regras são apresentados na Figura 4 para as medidas MAD, desvio padrão, 6 pseudo-classes e 6 intervalos do Mineset.

Tabela 7: Valores para as diferentes medidas de erro para regras da Base de Dados Housing.

		Ajustamento			Pseudo-Classe		Discretização Baseada em Entropia		
Número Regra	MAD	Desvio Padrão	DP e Média	DP e Mediana	3 pseudo-classes	6 pseudo-classes	3 intervalos Mineset	6 intervalos Mineset	Mineset automático
R0001	0,18	0,01	0,00	0,00	0,05	0,28	0,30	0,48	0,50
R0002	0,00	0,51	0,13	0,17	0,05	0,21	0,19	0,36	0,49
R0003	1,00	0,60	0,20	0,20	0,20	0,40	0,20	0,60	0,60
R0004	0,57	0,28	0,02	0,05	0,10	0,28	0,01	0,06	0,68

Já na Figura 7 os gráficos *boxplot* representam a diferença entre a medida MAD e todas as estratégia utilizadas para a Base de Dados Housing. Para essa base, nada pode ser concluído, pois a hipótese H_0 foi rejeitada para todas as estratégias, ou seja, existe diferença significativa entre as medidas. Nos gráficos isso é mostrado pelo fato de H_0 não estar contido no intervalo de confiança.

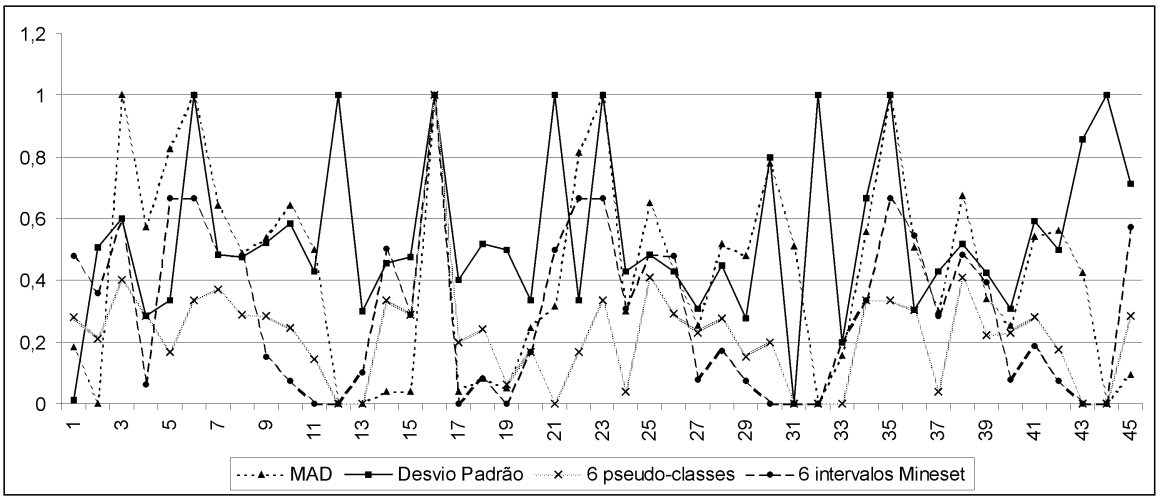
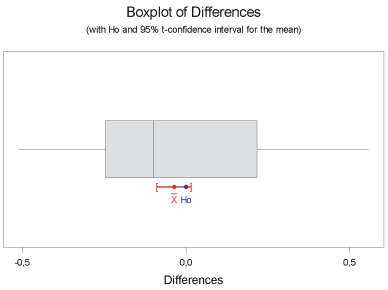
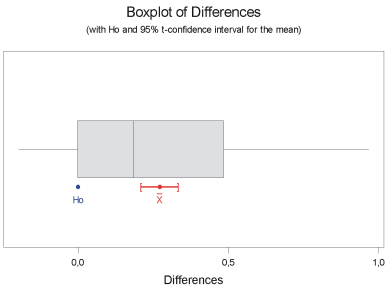


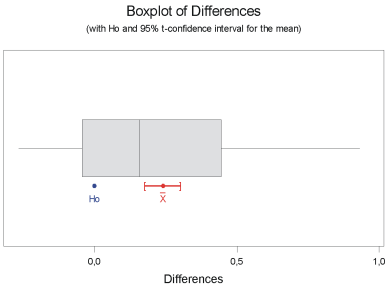
Figura 4: Gráfico de quatro medidas para as regras da Base de Dados Housing.



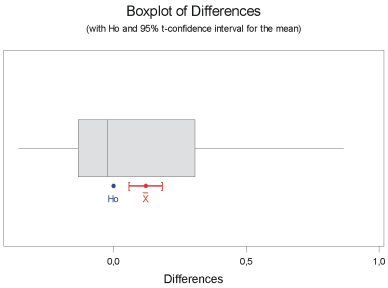
(a) Desvio Padrão



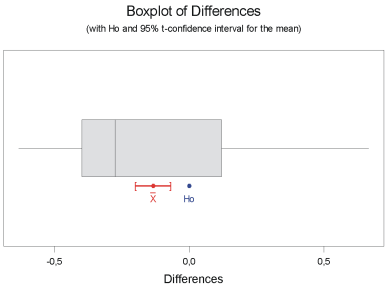
(b) Desvio Padrão e Média



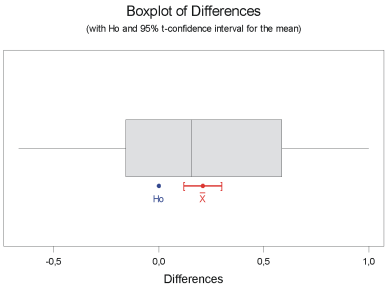
(c) Desvio Padrão e Mediana



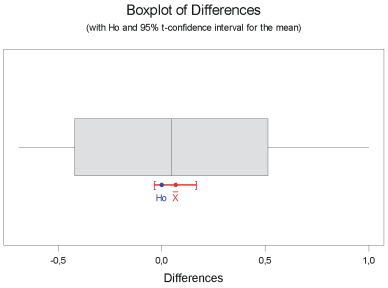
(d) 3 Pseudo-classes



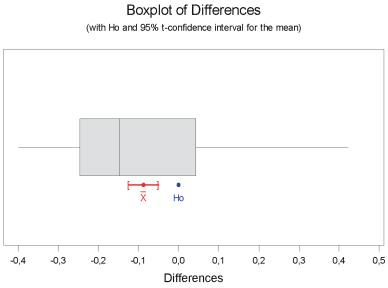
(e) 6 Pseudo-classes



(f) 3 Intervalos Mineset

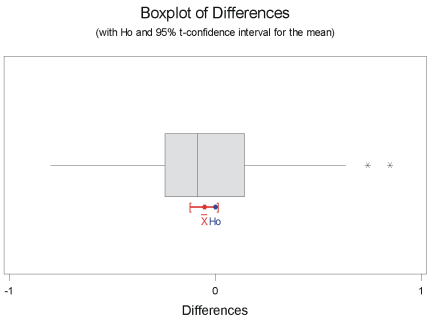


(g) 6 Intervalos Mineset

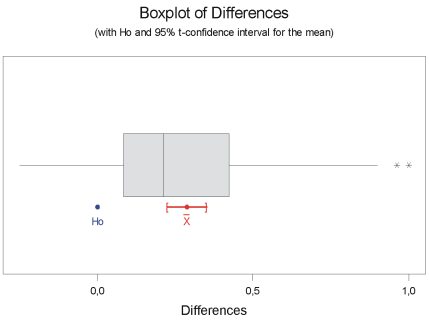


(h) Mineset Automático

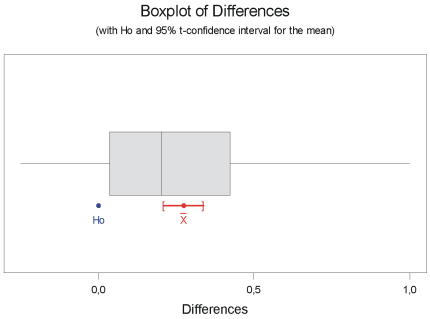
Figura 5: Boxplot da Base de Dados Abalone.



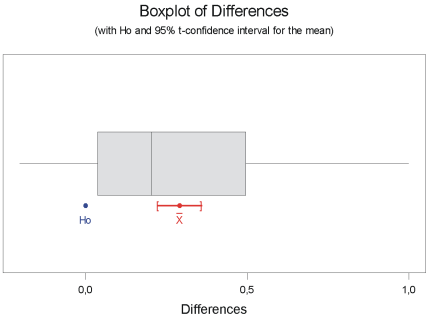
(a) Desvio Padrão



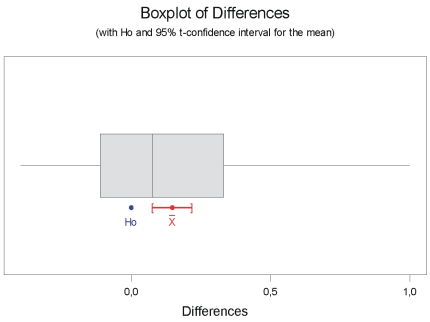
(b) Desvio Padrão e Média



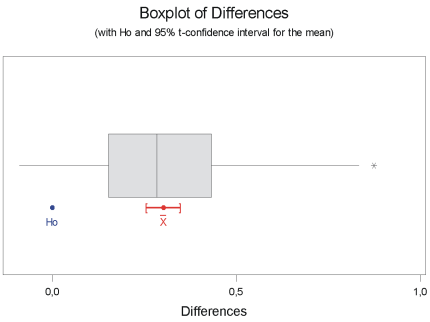
(c) Desvio Padrão e Mediana



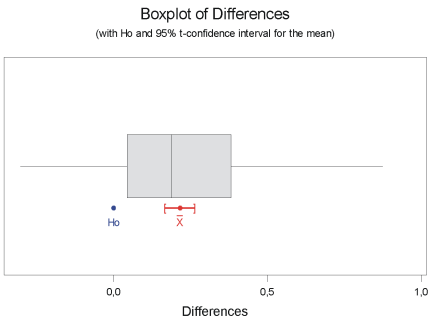
(d) 3 Pseudo-classes



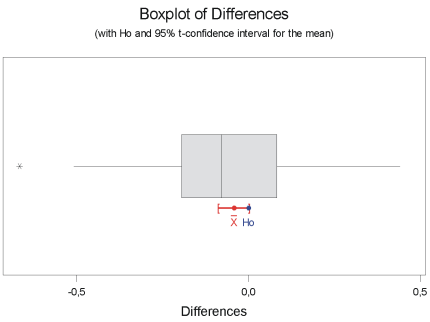
(e) 6 Pseudo-classes



(f) 3 Intervalos Mineset

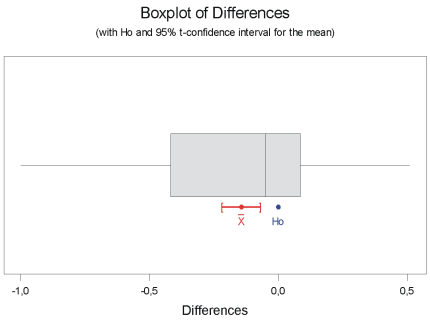


(g) 6 Intervalos Mineset

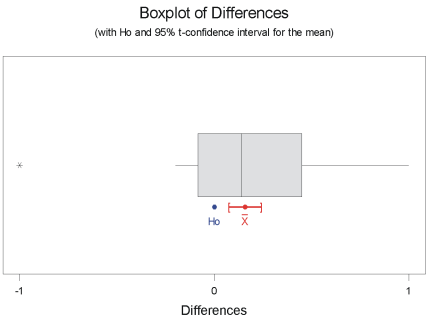


(h) Mineset Automático

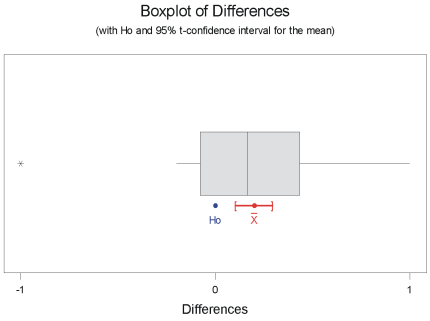
Figura 6: Boxplot da Base de Dados Auto-Mpg.



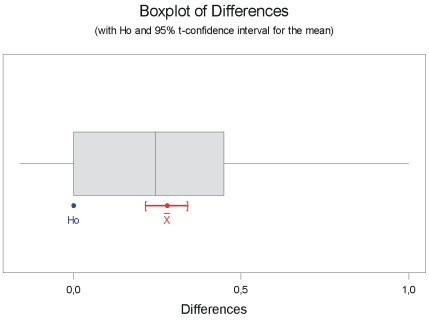
(a) Desvio Padrão



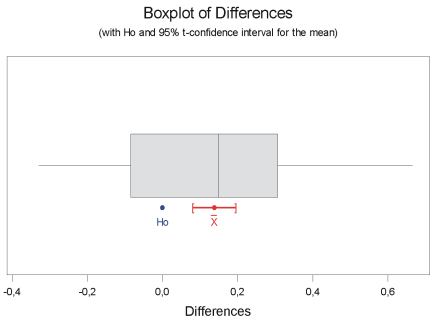
(b) Desvio Padrão e Média



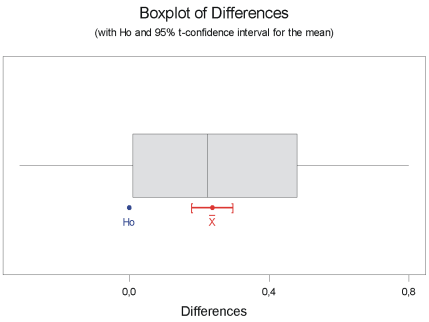
(c) Desvio Padrão e Mediana



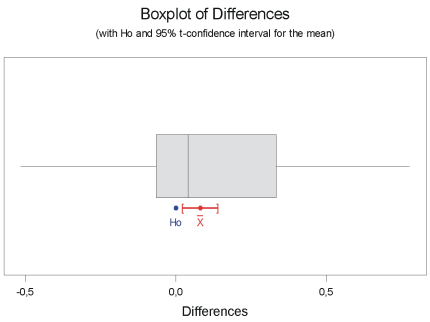
(d) 3 Pseudo-classes



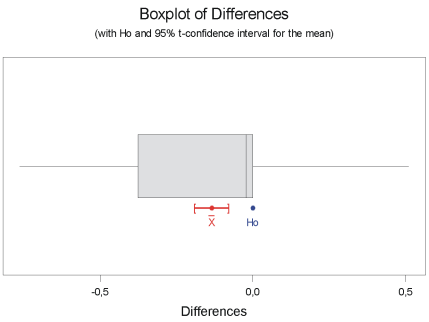
(e) 6 Pseudo-classes



(f) 3 Intervalos Mineset



(g) 6 Intervalos Mineset



(h) Mineset Automático

Figura 7: Boxplot da Base de Dados Housing.

5 Considerações Finais

Entre os tópicos de Mineração de Dados que têm recebido grande atenção por diversos pesquisadores estão a avaliação tanto do desempenho quanto da qualidade do conhecimento obtido.

Se for qualificada como conhecimento, a saída de um algoritmo de aprendizado deve ser precisa, estável e compreensível. Uma das formas de se obter diversas medidas para avaliação de conhecimento é através da matriz de contingência. Porém, para problemas de regressão, a construção dessa matriz não é trivial por ser o atributo meta contínuo. Visando solucionar o problema da matriz de contingência para regressão, neste relatório técnico foram experimentadas três estratégias para construção dessa matriz. Em cada estratégia também foram consideradas algumas variações.

Após a execução de todos os experimentos para as três Bases de Dados, pode-se verificar que, comparando-se com a medida MAD, a única medida que não rejeitou a hipótese H_0, ou seja, não houve diferença significativa entre elas em duas das três Bases de Dados foi a que utilizou a estratégia de ajustamento considerando somente o desvio padrão.

Assim, a matriz de contingência calculada usando a estratégia de ajustamento aqui proposta leva a um comportamento semelhante ao de uma medida consagrada para avaliar o erro de regras de regressão, que é a medida MAD. Ressalta-se a necessidade de realizar mais experimentos para melhor validação dos resultados.

Com isso, há um aumento no número de medidas que podem ser usadas para avaliar regras de regressão, uma vez que todas as já definidas e derivadas da matriz de contingência podem então ser utilizadas.

Referências

Allen, A. O. (1990). *Probability, Statistics, and Queueing Theory with Computer Science Applications*. Academic Press, Inc.

Blake, C. L. & C. J. Merz (1998). UCI repository of machine learning databases. Disponível em: `http://www.ics.uci.edu/~mlearn/MLRepository.html`.

Cheng, S. (1998, Feb). Statistical approaches to predictive modeling in large databases. Dissertação de Mestrado, Simon Fraser University, British Columbia, Canada.

Clark, P. & T. Niblett (1989). The CN2 Induction Algorithm. *Machine Learning 3*(4), 261–283. Disponível em `http://citeseer.nj.nec.com/clark89cn.html`. Acesso em 20/03/2003.

Dosualdo, D. G. (2003). Investigação de regressão no processo de mineração de dados. Dissertação de Mestrado, Instituto de Ciências Matemáticas e de Computação - USP São Carlos.

Lavrač, N., P. Flach, & R. Zupan (1999, jun.). Rule evaluation measures: A unifying view. In S. Dzeroski & P. Flach (Eds.), *Proceedings of the Ninth International Workshop on Inductive Logic Programming (ILP-99)*, Volume 1634, pp. 74–185. Springer-Verlag. Lecture Notes in Artificial Intelligence.

Liu, B. & W. Hsu (1996). Post-analysis of learned rules. *AAAI*, 828–834.

Liu, B., W. Hsu, & S. Chen (1997). Using general impressions to analyse discovered classification rules. *American Association for Artificial Intelligence*.

Magalhães, M. N. & A. C. P. de Lima (2002). *Noções de Probabilidade e Estatística*. Edusp — Editora da Universidade de São Paulo.

Merz, C. J. (1998). *Classification and Regression by Combining Models*. Tese de Doutorado, University of California, Irvine. Disponível em: `http://www.ics.uci.edu:80/~cmerz/thesis.ps`.

Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill Series in Computer Science.

Pugliesi, J. B., D. G. Dosualdo, & S. O. Rezende (2003). Sintaxe padrão para representar regras de regressão no discover. Technical Report 193, ICMC – USP, São Carlos. Available: `ftp://ftp.icmc.usp.br/pub/BIBLIOTECA/rel_tec/RT_193.zip`.

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann.

Rathjens, D. (1996). MineSet^{TM} *User's Guide*. Silicon Graphics, Inc.

Rulequest-Research (2000, 03). Rulequest research – Data mining tools. Disponível em: `http://www.rulequest.com/cubist-info.html`.

Torgo, L. (1995). Data fitting with rule-based regression. In *Proceedings of the 2nd International Workshop on Artificial Intelligence Techniques (AIT'95)*, Brno, Czech Republic, pp. 223–232. J. Zizka & P. Brazdil.

Torgo, L. F. R. A. (1999, Sep). *Inductive Learning of Tree-Based Regression Models*. Tese de Doutorado, Faculdade de Ciências da Universidade do Porto. Disponível em: `http://www.ncc.up.pt/~ltorgo/PhD/`.

Weiss, S. M. & N. Indurkhya (1995). Rule-based machine learning methods for functional prediction. *Journal of Artificial Intelligence Research 3*, 383–403. Disponível em: `http://www.cs.cmu.edu/afs/cs/project/jair/pub/volume3/weiss95a.ps`.