3
Metodologia

Neste capítulo apresentar-se-á a metodologia estruturada e métodos utilizados nesta dissertação para a identificação dos atributos mais relevantes para a determinação do desfecho (grau de investimento ou especulação de empresas), e a posterior predição deste desfecho.

Esta metodologia é constituída por duas macro etapas, a saber:

Etapa 1: Seleção de atributos;
Etapa 2: Predição do desfecho.

Na seção 3.1 descreve-se os métodos utilizados para a seleção de atributos (Etapa 1) e na seção 3.2 os métodos utilizados para a predição do desfecho (Etapa 2).

3.1
Seleção de atributos

Na presente aplicação, a utilização de métodos de seleção de atributos visa dois objetivos centrais: (i) determinar, entre os atributos normalmente medidos, quais são de fato importantes na estimação do *rating* (grau) e (ii) melhorar o desempenho, no sentido de estimar o *rating* (grau) de forma automática, sem componentes subjetivos. Serão explorados, e comparados os resultados de métodos lineares e não-lineares:

- **Lineares**: Correlação, *Stepwise* e Análise de Componentes Principais (PCA);
- **Não-lineares**: Informação Mútua (IM) e MIFS-U (um procedimento baseado em IM);

Na seqüência descreve-se brevemente cada um destes procedimentos.

3.1.1
Correlação

Sejam x_1 e x_2 duas variáveis aleatórias (V.A.'s). Podemos definir o coeficiente de correlação *'ρ'* como:

$$\rho(x_1, x_2) \equiv \frac{Cov(x_1, x_2)}{\sigma_{x_1}\sigma_{x_2}} = E\left[\left(\frac{x_1 - E(x_1)}{\sigma_{x_1}}\right)\left(\frac{x_2 - E(x_2)}{\sigma_{x_2}}\right)\right]$$

Na presente aplicação estas V.A.'s descreverão possíveis atributos ou o desfecho (aqui denotado como '*y*'). O coeficiente de correlação $\rho(x,y)$ representa a dependência linear de duas V.A.'s .

Assim um atributo, x, e o desfecho, y, podem ser descritos como:

- positivamente correlatados se $\rho(x,y) > 0$
- negativamente correlatados se $\rho(x,y) < 0$
- descorrelatados se $\rho(x,y) = 0$

Como é bem conhecido, o fato de x e y serem independentes implica em x e y serem descorrelatados, porém se x e y forem descorrelatados não necessariamente implica em x e y serem independentes (James, 2002).

3.1.2
Stepwise

Se assumirmos que o desfecho, y, guarda uma relação linear com os atributos $x_1, x_2, ..., x_m$, pode-se utilizar o procedimento conhecido como *Stepwise* (McClave, 2005) brevemente descrito a seguir*:*

Passo 1: Para todos os possíveis modelos lineares da forma:

$$E(y) = \beta_0 + \beta_1 x_i$$

Onde x_i é um atributo independente e *i=1,..., m.*

Formula-se um teste de hipótese sendo a hipótese nula:

$$H_0 : \beta_1 = 0$$

e a hipótese alternativa:

$$H_a : \beta_1 \neq 0$$

Usa-se o teste *t* para um único parâmetro β. O atributo independente x_i que produzir o maior valor absoluto *t* é declarado o melhor atributo preditor de y.

Passo2: De posse desta informação, o método *Stepwise* irá procurar então entre os m-1 atributos restantes, o melhor modelo de dois atributos da forma:

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_i$$

Este procedimento é feito comparando todos os modelos de dois atributos independentes que contém x_1 e cada uma das *(m-1)* opções para o segundo atributo x_i. Os valores dos testes *t* H_0: $\beta_2=0$ são calculados para cada *(m-1)* modelos e o atributo que tem o maior *t* é escolhido e definido como atributo x_2. O algoritmo agora compara o valor do teste *t* de $\hat{\beta}_1$ após $\hat{\beta}_2 x_2$ ter sido adicionado pelo modelo. Se o valor *t* se tornou não significativo em algum especificado nível α , o atributo x_1 é removido e uma busca é realizada para achar o atributo independente com um parâmetro β que manterá o valor *t* mais significante dada a presença de $\hat{\beta}_2 x_2$.

O Passo 2 é realizado novamente, para as *(m-2)* opções restantes e assim sucessivamente até que o modelo só contenha os atributos com valores *t* que sejam significantes dado um nível α. Assim, na maioria das situações práticas, apenas alguns dos atributos do conjunto *m* estarão presentes no modelo.

3.1.3
Análise de Componentes Principais (PCA)

A análise de componentes principais (*PCA – Principal Component Analysis*) (Dunteman, 1989; Haykin, 1998) é um dos métodos mais populares em estatística multivariada e pode ser usado para compressão de dados ou seleção de atributos. Este método gera um conjunto novo de atributos, chamado componentes principais. Cada componente é uma combinação linear dos atributos originais,

sendo que todas estas componentes são ortogonais entre si, de maneira que não haja informação redundante. As componentes principais com um todo, formam uma base ortogonal no espaço dos dados.

A primeira componente principal pode ser vista como um eixo no espaço, que possui variância máxima entre todos os possíveis eixos. A segunda componente principal é ortogonal à primeira, sendo a sua variância máxima entre todas as possíveis escolhas para este segundo eixo. É possível então 'truncar' este processo de geração de componentes quando a variância já estiver suficientemente representada. Nesta dissertação consideramos uma meta de 95% da variância total acumulada da(s) primeira(s) componente(s) para considerar que a variância dos dados originais está bem representada por estas componentes. Para tal podemos, por exemplo, considerar a importância dos atributos em cada componente de forma ponderada, utilizando a equação a seguir:

*relevância_relativa = relevância * variância_percentual/100;*

onde:

relevância_relativa é a relevância do atributo ponderada pela importância da componente principal;
relevância é a relevância do atributo apenas nesta componente principal;
variância_percentual é o quanto percentualmente esta componente principal participa da variância total original dos dados.

Ponderando cada um dos atributos pelas importâncias relativas em cada componente, podemos posteriormente ordená-los de acordo com suas relevâncias em relação ao desfecho.

Entretanto, a maior desvantagem do PCA está em não ser invariante à transformações nos atributos, ou seja, uma simples mudança de métrica é suficiente para modificar seus resultados.

3.1.4 Informação Mútua

Durante o processo de seleção de atributos, deseja-se identificar aqueles atributos que são importantes, isto é, com muita informação relativa à saída. Para

tal, é necessário medir essa informação. A teoria da informação fornece um método para medir a informação de variáveis aleatórias: a entropia e a informação mútua (Shannon *et al.*, 1949; Cover *et al.*, 1991). Sejam '*x*' e '*y*' variáveis aleatórias contínuas com função de densidade de probabilidade (pdf) *p(x)*. A entropia de *x*, representada como *H(X)* e a informação mútua entre *x* e *y*, representada como *I(X,Y)* são definidas como:

$$H(X) = -\int p(x) \log p(x) dx$$

$$I(X,Y) = \int p(x,y) \log \frac{p(x,y)}{p(x)p(y)} dxdy$$

A entropia é uma medida de incerteza de variáveis aleatórias, e a informação mútua é uma maneira de medir o quanto de informação a variável aleatória *x* possui em relação a variável aleatória *y*. Se a informação mútua entre duas variáveis aleatórias é grande (pequena), significa que as duas variáveis são muito (pouco) relacionadas. Se a informação mútua é próxima de zero, as duas variáveis aleatórias são independentes.

Considerando-se os atributos e o desfecho como variáveis aleatórias podemos caracterizar a relevância entre os atributos perante o desfecho em termos da correlação ou da informação mútua (Peng et al., 2005). No entanto, métodos baseados na dependência linear de atributos, como a correlação, por exemplo, podem não conseguir capturar relações arbitrárias entre padrões de comportamento de atributos e suas diferentes classes. A informação mútua apresenta ainda uma vantagem importante que é a independência de coordenadas, ou seja, não é afetada pela normalização dos dados, o que contribui para a robustez do método.

3.1.5 MIFS-U

No processo de selecionar atributos, reduzir o seu número, de maneira que sejam excluídos atributos irrelevantes ou redundantes é útil na maximização do

desempenho de modelos preditores. Logo, é desejável descobrir um conjunto de atributos que possua a máxima relevância possível (medido através da informação mútua) com o desfecho, e a mínima redundância possível entre os atributos. Porém, para determinar este conjunto ótimo de atributos de maneira exaustiva seria necessário, por exemplo, calcular todas as informações mútuas entre todos os atributos de todas as possíveis combinações de todo o conjunto de atributos, o que implicaria no cálculo de todas as funções de densidade de probabilidade conjuntas de todas as combinações possíveis entre os atributos e o desfecho.

Para contornar este problema, (Kwak, 2002) propõe o algoritmo de Seleção de Variáveis sob Informação Mútua com Distribuição Uniforme de Informação (MIFS-U). Este algoritmo busca ordenar os atributos de acordo com suas relevâncias (medidas através da informação mútua) em relação ao desfecho considerando a informação mútua entre os atributos, e é descrito brevemente a seguir:

Seja F o conjunto de *m* atributos candidatos f_i *onde i=1,...,m* para o conjunto de atributos relevantes e C o conjunto que representa o desfecho (grau de investimento ou especulação).

Passo 1 – Inicialize um conjunto S = ∅ e calcule a Informação Mútua entre os atributos e o desfecho: $I(C; f_i)\ \forall\ f_i \in F$.

Passo 2 – Selecione o primeiro atributo considerando-o como o que maximiza $I(C; f_i)\ \forall\ f_i \in F$, e faça $F \leftarrow F - \{f_k\}$ e $S \leftarrow f_k$

Passo 3 – Escolha $f_k \in F$ de maneira que $I(C; f_i, S)$ é maximizada para $f_i \in F$.

Porém, o cálculo de $I(C; f_i, S)$ no passo 3 não é trivial (Kwak, 2002), porém pode-se considerar:

$$I(C; f_i, f_S) = I(C; f_S) + I(C; f_i \mid f_S) \quad \textit{para um dado } f_S \in S.$$

Note que $I(C; f_S)$ é igual para todos os atributos candidatos, logo seu cálculo não é necessário no processo de maximização. Desta maneira, a

Informação mútua condicional pode ser representada (Kwak, 2002) como:

$$I(C; f_i \mid f_S) = I(C; f_i) - \{ I(f_S; f_i) - I(f_S; f_i \mid C)\}.$$

Que pode ser reescrita como:

$$I(C; f_i \mid f_S) = I(C; f_i) - (I(C; f_S) / H(f_S))\, I(f_S; f_i)$$

Onde $H(f_S)$ representa a entropia para um dado $f_S \in S$. Um procedimento de maximização iterativo pode ser implementado calculando a entropia $H(f_S)$ e $I(f_S; f_i)$.

3.2
Predição do desfecho

Nesta seção, o principal objetivo é a predição do desfecho (grau de investimento ou especulação da empresa) a partir de um conjunto de atributos previamente selecionados. Iremos considerar métodos:

- **Lineares**: Regressão Múltipla Linear e Discriminante Linear de Fisher e;
- **Não Lineares**: Redes Neurais *Multi Layer Perceptrons*.

Classificadores tendem a ter a sua taxa de acerto maximizada quando os conjuntos de dados de aprendizado são balanceados. Neste presente estudo, balanceou-se os dados disponíveis multiplicando o conjunto de observações de grau de especulação (inicialmente com 56 empresas), por cinco, resultando em um total de 280 observações para esta classe. Manteve-se o número de empresas classificadas como grau de investimento, contabilizando 262 observações, totalizando o conjunto de 542 observações de ambos os graus, utilizadas para a fase de treinamento. Desta maneira, o conjunto de dados utilizados no treinamento é composto por 51.66% de observações da classe de grau de especulação e 48.34% de observações da classe de grau de investimento.

Sabe-se que atributos com grandes diferenças em sua magnitude, podem comprometer o desempenho do poder classificatório, já que durante a fase de aprendizado (ou ajuste de parâmetros) haveria uma ponderação artificialmente

criada no erro produzido por parte das observações. Normalizamos os dados dos atributos de entrada no intervalo [-1,1] utilizando a função:

$$pn = 2*(p-minp)/(maxp-minp) - 1$$

onde:

pn = valor normalizado;

minp = menor valor contido no atributo

maxp = maior valor contido no atributo

p = valor não-normalizado

Todos os métodos de predição utilizados neste presente estudo fizeram uso do banco de dados normalizado e balanceado.

3.2.1
Métodos de classificação lineares

3.2.1.1
Regressão Múltipla Linear

Suponha que o seguinte modelo linear é valido para relacionar o desfecho y e os atributos $x_1, x_2,..., x_m$, com erros normais ε:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_m x_{i,m} + \varepsilon_i$$

onde,

$\beta_0, \beta_1,..., \beta_m$, são os parâmetros a serem estimados ;

m é o número de parâmetros independentes;

$x_{i1}, x_{i2},..., x_{im}$ são atributos mensurados;

e ε_i i=1,2,...,n são erros independentes com distribuição N(0, σ^2)

n é o número de observações disponíveis.

Então o valor esperado do desfecho é dado por:

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_m x_m$$

Nota-se que neste modelo, o efeito de x_1 sobre a resposta média não depende de x_2 e vice-versa, assim os atributos preditores têm efeito aditivo ou não interagem. Para o cálculo dos valores dos parâmetros $\beta_0, \beta_1, ..., \beta_m$, utilizamos o método de mínimos quadrados (McClave, 2005).

3.2.1.2 Discriminante Linear de Fisher

O método de Fisher (Duda, 2001) busca a combinação linear que maximiza a razão da distância ao quadrado entre as projeções dos centróides dos dois grupos para a variância das projeções. A saber, um centróide é um ponto do espaço p-dimensional cujas coordenadas são as médias aritméticas dos atributos discriminantes para os indivíduos pertencentes ao mesmo grupo.

Seja b o vetor com os pesos da combinação linear. Então, $y_{(i)}$, o valor da projeção de $X_{(i)}$ sobre b, é dado por:

$$y_{(i)} = b^T X_{(i)}$$

Sejam $\overline{X_1}$ e $\overline{X_2}$ os centróides dos grupos 1 e 2, respectivamente. Então, as médias das projeções sobre b são dadas por:

$$\overline{y_1} = b^T \overline{x_1} \text{ e } \overline{y_2} = b^T \overline{x_2}$$

Supondo que as matrizes de covariância de ambos os grupos são iguais e denotadas por C_x, tem-se que a variância das projeções para qualquer dos grupos é dada por:

$$S_y^2 = b^T C_x$$

O objetivo do método é encontrar b que maximiza a proporção entre a diferença das médias de $\overline{y_1}$ e $\overline{y_2}$ ao quadrado a variância de y, dada por:

$$\Delta = \frac{(b^T(\overline{x_1} - \overline{x_2}))^2}{b^T C_x b}$$

A solução deste problema é dada por:

$$b = S^{-1}(\overline{x_1} - \overline{x_2})$$

onde *S* é a matriz de covariância (comum aos dois grupos) estimada por:

$$S = \frac{1}{n_1 + n_2 - 2}(x_1^T x_1 + x_2^T x_2)$$

e onde x_1 e x_2 são as matrizes de dados referentes aos grupos 1 e 2, com respectivamente, n_1 e n_2 indivíduos.

A chamada função discriminante de Fisher é dada por:

$$Y_{(i)} = b^T X_{(i)} = (\overline{x_1} - \overline{x_2})^T S^{-1} X_{(i)}$$

A função discriminante de Fisher pode ser usada para construir uma regra de decisão: uma observação $X_{(i)}$ será classificada no grupo cuja média está mais próxima. Isto é, se $\left|y_{(i)} - \overline{y_1}\right| < \left|y_{(i)} - \overline{y_2}\right|$ então $X_{(i)}$ é classificado no grupo 1.

3.2.2
Método de classificação não-linear

3.2.2.1
Redes Neurais Artificiais

As Redes Neurais Artificiais (RNAs) são uma metodologia cujas raízes abrangem neurociência, matemática, estatística, física, ciência da computação e engenharia. Segundo (Haykin, 1998), uma Rede Neural é um processador maciçamente paralelamente distribuído constituído de unidades de processamento simples (denominados neurônios), que tem a propensão natural para armazenar conhecimento experimental e torná-lo disponível para o uso.

A motivação original desta metodologia foi a de modelar a rede de neurônios humanos visando compreender e reproduzir o funcionamento do cérebro humano. Uma RNA se assemelha ao cérebro humano em dois aspectos:

1. O Conhecimento é adquirido pela rede a partir de seu ambiente através de um processo de aprendizagem;

2. Forças de conexão entre neurônios, conhecidas como pesos sinápticos, são utilizadas para armazenar o conhecimento adquirido.

O procedimento utilizado para realizar o processo de aprendizagem ou treinamento é chamado de algoritmo de aprendizagem, cuja função é modificar os pesos sinápticos da rede de uma forma ordenada para alcançar um objetivo de projeto desejado.

A utilização de Redes Neurais oferece, dentre outras as propriedades de não-linearidade, mapeamento de entrada-saída, adaptabilidade e robustez.

Nesta dissertação, utilizou-se uma Rede Neural *feedforward* tradicional com algoritmos de retro-propagação de erros (*Back-Propagation*) com Regularização Bayesiana (Haykin, 1998) para treinamento. A arquitetura inicial é composta por dez neurônios em uma camada escondida e um único neurônio na camada de saída. Em todos os neurônios (sejam da camada escondida ou da camada de saída) utilizou-se a função de ativação sigmóide.

O algoritmo de Regularização Bayesiana (RB), adiciona um termo de penalização (regularização) à função objetivo, de forma que o algoritmo de estimação faça com que os parâmetros irrelevantes convirjam para zero, reduzindo assim o número de parâmetros efetivos utilizados no processo. O algoritmo de RB assume que os parâmetros da rede são variáveis aleatórias com distribuições especificadas. Os parâmetros de regularização são variâncias desconhecidas associadas a estas distribuições e pode-se calcular estes parâmetros usando técnicas estatísticas. Com este procedimento, o modelo de RNA (por exemplo, número de neurônios) utilizado não é especificado de maneira empírica e arbitrária.

Devido ao pequeno número de observações disponíveis no banco de dados utilizou-se para avaliar a capacidade de generalização dos procedimentos o "método deixe um de fora" (*leave-one-out*; Haykin, 1998; Bishop, 1995). Neste método, *N-1* observações são utilizados para treinar o modelo, que é validado testando-o como exemplo deixado de fora. O procedimento é realizado novamente *N* vezes, de maneira que em cada rodada, uma observação diferente é deixada para a validação. O erro quadrado na validação é a média sobre as *N* validações do procedimento. O resultado de generalização é uma ótima aproximação do erro fora-da-amostra.