UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"

Ralph Pinotti Leite

Estratégia de Investimento baseada em padrões descobertos por mineração de dados em séries históricas de preços diários do Índice IBOVESPA e das ações da PETROBRAS-PN (PETR4.SA)

Botucatu – SP

2010

Ralph Pinotti Leite

Estratégia de Investimento baseada em padrões descobertos por mineração de dados em séries históricas de preços diários do Índice IBOVESPA e da ações da PETROBRAS-PN (PETR4.SA)

Monografia apresentada ao Instituto de Biociências da Universidade Estadual Paulista "Júlio de Mesquita Filho", Campus de Botucatu, para a obtenção do título de Bacharel em Física Médica.

Orientador:
Prof. Dr. Ney Lemke

Bacharelado em Física Médica
Departamento de Física e Biofísica
Instituto de Biociências
Universidade Estadual Paulista "Júlio de Mesquita Filho"
Campus de Botucatu

Botucatu – SP

2010

FICHA CATALOGRÁFICA ELABORADA PELA SEÇÃO TÉC. AQUIS. E TRAT. DA INFORMAÇÃO
DIVISÃO TÉCNICA DE BIBLIOTECA E DOCUMENTAÇÃO - CAMPUS DE BOTUCATU - UNESP
BIBLIOTECÁRIA RESPONSÁVEL: ***ROSEMEIRE APARECIDA VICENTE***

Pinotti, Ralph Leite.
Estratégia de Investimento baseada em padrões descobertos por mineração de dados em séries históricas de preços diários do Índice IBOVESPA e das ações da PETROBRAS-PN (PETR4.SA) / Ralph Pinotti Leite. - Botucatu, 2010

Trabalho de conclusão de curso (bacharelado - Física Médica) - Instituto de Biociências de Botucatu, Universidade Estadual Paulista, 2010
Orientador: Ney Lemke
Capes: 10501002

1. Mineração de dados (Computação). 2. Algoritmos. 3. Investimento – Projetos de.

Palavras-chave: Algoritmo J4.8; Estratégia de investimento; Mineração de dados.

À minha mãe e irmã, Regina Marcia Pinotti e Priscila Pinotti Leite,
exemplos de honestidade e esforço,
que tornaram possível esta conquista.

Agradecimentos

Meus mais sinceros agradecimentos à todos que me ajudaram na elaboração desse trabalho:

– Ao Professor Doutor Ney Lemke, pela orientação e incentivo;

– À equipe do Laboratório de Bioinformática e Biofísica Computacional do Departamento de Física e Biofísica do IBB-Unesp, em especial ao doutorando Marcio Luis Acencio, pela ajuda e incentivo, ao amigo Luiz Augusto Bovolenha pela indispensável ajuda na área computacional e ao amigo Pedro Rafael Costa pelas grandes discussões onde muitas idéias eram criadas;

– À minha mãe, Regina Marcia Pinotti e a minha irmã, Priscila Pinotti leite pela compreensão, auxilio e disposição em qualquer momento que eu precisa-se;

– À IV Turma de Física Médica da Unesp de Botucatu, com os quais dividi momentos de alegria e de desespero durante os mais de 4 anos de graduação;

– À todos os verdadeiros amigos que foram feitos nas repúblicas por onde passei, onde foram minha família na permanência na cidade de Botucatu e por fazerem parte da formação do meu caráter;

– À todos os amigos da IBBJr. (Empresa Junior do instituto de Biociência da UNESP -Botucatu), por todo o aprendizado, espírito de trabalho em equipe que consegui;

Resumo

Com uma quantidade cada vez maior de investidores e consequentemente maior número de transações feitas, o estudo de novas estratégias de investimento na bolsa de valores com técnicas de mineração de dados vem sendo foco de crescente interesse em pesquisas. Estas permitem que uma quantidade de dados históricos sejam processados e analisados visando descobrir termos e padrões que possam ser úteis na tomada de decisão de um investimento. Visando obter lucro com aplicações acima do desempenho real dos índices analisados, propomos neste trabalho, um método de estratégia de investimento utilizando regras geradas por algoritmos classificadores. Para isso, os dados históricos diários do índice IBOVESPA e das ações da Petrobras (PETR4.SA) são organizados e processados determinando quais os principais atributos que influenciam o índice decisivamente quando toma-se uma decisão de investimento. Para mostrar a validade das regras, carteiras de investimento fictícias são elaboradas, mostrando o desempenho das decisões perante o desempenho real do índice e das ações. Os resultados mostram que a estratégia de investir segundo a regra gerada, retorna ganho superior que o real desempenho da Bolsa de Valores. A característica de cada classificador maximiza o ganho no período analisado permitindo inferir o retorno que essa técnica pode dar e quanto tempo leva para dobrar o valor inicial investido. O melhor classificador aplicado sobre a série histórica e seu uso na estratégia de investimento proposta demandaria 104 dias para dobrar o investimento inicial.

Palavras-chave: Mineração de dados, Algoritmo J4.8,estratégia de investimento.

Abstract

With the increase of stakeholders and consequently increase of amount of financial transaction the study of news investment strategies in the stock market with data mining techniques has been the target of important researches. It allows that great historical data base to be processed and analysed looking for pattern that can be used to take a decision in investments. With the idea of getting profit more than the real indexs' gain, we propose a strategy method of transactions using rules built by algorithm classification. For that, diary historical data of Ibovespa index and Petrobras stocks are organized and processed to finding the most important attribute that act decisively when taking a investment decision.To test the accuracy of proposed rules, a non real portfolio management is created, showing the decisions' performance over the real index and stocks' performance. Following the proposed rules, the results show that the strategy of investment give me back a high return that Stock market's return. The exclusive characteristics of algorithms maximize the gain inside the analysed time allowing to determine the techniques' return and the number of the days necessary to double the initial investment. The best classifier applied on the time series and its use on the propose investments strategy will demand 104 days to double the initial capital.

Keywords: Data mining, Algorithm J4.8; investment strategy.

Sumário

Lista de Figuras

Lista de Tabelas

1 Introdução

1.1 Econofísica

Econofísica usa os conceitos da física estatística e física teórica na descrição de sistemas financeiros. Esses conceitos são aplicados à séries temporais de dados financeiros para obter *insights* sobre comportamento do mercado (MANTEGNA, R.; STANLEY, H;, 2000).

A econofísica tem por objetivo prever bolhas financeiras, calcular riscos e determinar preços de derivativos, entre outros. Econofísicos sugerem que a economia funciona como um grande sistema complexo guiado por alguma lei.

Todas essas teorias se desenvolveram em um campo interdisciplinar onde a finalidade é as vezes diferente da de um economista. Esses não a consideram complexa e tende a se preparar para lidar quando a bolha explode, outros tentam descobrir qual o momento mais propício em que essa pode explodir.

Algumas estratégias de investimento baseada em padrões encontrados em base de dados já foram testadas e comprovadas que funcionam, mas não podem dizer que tem poder preditivo.

Apesar de grandes idéias já terem sido derrubadas pelo próprio mercado, uma leva imensa de pesquisas, teorias e quantidade cada vez maior de profissionais usando ferramentas da ciências naturais na área, dá credibilidade a essa linha de pesquisa que substitui e complementa antigas teorias.

1.1.1 Previsões de séries temporais

É muito difícil fazer previsões de séries temporais financeiras, não pela falta de informações, mas sim pela enorme quantidade de dados que existem para serem estudados (Swingler, K. 1994).

Uma abordagem simplificada é delimitar padrões nas séries históricas, quando sabemos o que procede a esse e testar em dados mais recentes. Quando os padrões se repetem várias vezes, podemos confiar de que estão de alguma forma, relacionados.

Um algoritmo de predição pode ser feito usando métodos estatísticos, aprendizagem de máquina e análise dinâmica de um sistema, juntos fica conhecido como Mineração de dados.

Concentrado em extrair informações de dados crus. A predição tem que ser feita cuidadosamente para ver se preenche alguns critérios necessários e também é um procedimento que envolve uma quantidade de passos pré determinado que influencie na qualidade dos resultados (Zemke, S; 2003).

Para ser efetivo, um sistema preditivo precisa ter um bom conjunto de dados, uma ótima habilidade em descobrir e localizar parâmetro. O conjunto de dados precisa ser processado e técnicas de mineração de dados precisam ser bem empregadas. Comparar seu desempenho na bolsa de valores é uma maneira utilizada nesta tese para validar os resultados.

1.2 Mineração de Dados

Trata-se do processo de exploração de grandes quantidades de dados a procura de padrões consistentes, como regras de associações ou seqüências temporais, para detectar relacionamentos sistemáticos entre variáveis, detectando assim novos subconjuntos de dados (WITTEN; FRANK, 2000).

Para se obter as regras de associações para uma grande quantidade de dados, geralmente são utilizados algoritmos de aprendizagem de máquina, programas que melhoram seu desempenho por meio de experiências. São capazes de gerar hipóteses a partir dos dados, identificando padrões complexos que maximizam o índice de acerto da mineração.

O conceito de *Data Mining* está se tornando cada vez mais popular como uma ferramenta de gerenciamento de informação. Estas devem revelar estruturas de conhecimento, que possam guiar decisões em condições de certeza limitada.

Algumas dessas técnicas citadas são usadas em séries históricas financeiras a fim de encontrar padrões que possam ser usados para implementar estratégias de investimentos. Observando uma quantidade de dados de alguns índices e preços de ações, pode se inferir de certas movimentações que o mercado possa vir a ter.

1.3 Bolsa de valores

A BM&FBOVESPA é uma companhia de capital brasileiro formada, em 2008, a partir da integração das operações da Bolsa de Valores de São Paulo e da Bolsa de Mercadorias & Futuros. É a principal instituição brasileira de intermediação para operações do mercado de capitais. Os movimentos dos preços no mercado ou em uma seção dos mercados são capturados através de índices chamados Índice de Bolsa de Valores.

- Os preços das ações servem também para indicar o valor de mercado das empresas cotadas em bolsa.

O *after market* da Bovespa é um período do dia em que a bolsa de valores funciona após o pregão normal. Em linhas gerais, o after Bovespa nada mais é do que um horário extra de funcionamento da bolsa. Ele possibilita os investidores que não têm como acompanhar o mercado durante o dia, investir neste horário extra. Esse período é muito explorado por investidores pois, esses usam das informações do dia para gerenciar suas aplicações e ordens. Usar de técnicas computadorizadas e rápidas oferece certas vantagens entre aplicadores que buscam facilidades e precisão nas informações. Uns dos métodos de análise utilizados pelos investidores é a análise técnica. Essa análise busca padrões nas informações nos gráficos formados por informações dos preços das ações Esse tipo de informação quando bem interpretada, gera grandes indícios de como será o comportamento do ativo no dia seguinte. Analisando-se então alguma quantidade de dados, pode-se inferir a respeito da sua movimentação futura.

1.3.1 Ibovespa

Índice Bovespa é o mais importante indicador do desempenho médio das cotações do mercado de ações brasileiro. Sua relevância advém do fato do Ibovespa retratar o comportamento dos principais papéis negociados na BM&FBOVESPA

É o valor atual, em moeda corrente, de uma carteira teórica de ações constituída em 02/01/1968 a partir de uma aplicação hipotética. Supõe-se não ter sido efetuado nenhum investimento adicional desde então, considerando-se somente os ajustes efetuados em decorrência da distribuição de proventos pelas empresas emissoras. Dessa forma, o índice reflete não apenas as variações dos preços das ações, mas também o impacto da distribuição dos proventos, sendo considerado um indicador que avalia o retorno total de suas ações componentes.

A finalidade básica do Ibovespa é a de servir como indicador médio do comportamento do mercado. Para tanto, sua composição procura aproximar-se o mais possível da real configuração das negociações à vista na BM&FBOVESPA. É dentro desse cenário que o estudo comportamental do histórico desse índice pode influenciar em uma análise de qualquer ação que a compõe.

1.4 Estratégia de investimento

O índice Ibovespa e as ações da Petrobras contêm uma enorme quantidade de informações que são computadas a todo tempo. Essas informações são organizadas em forma de imensos bancos de dados que podem ser usados por pesquisadores, investidores, economista e qualquer um que se interessa em estudar seu comportamento. Com as técnicas de Mineração de Dados já descritas anteriormente, o objetivo é vasculhar minuciosamente por padrões já existentes nestes conjuntos de dados, com finalidade de serem usados como regra de decisão em estratégias de investimentos. Essa estratégia consiste em analisar vários grupos de dados contendo informações de 6 dias históricos consecutivos e com isso descobrir padrões nos 5 primeiros dias. Esses padrões organizados estatisticamente por relevância nos dizem de alguma forma o que pode ocorrer no dia 6.

Os modelos de regra de decisão quando gerados, classificam dados que não estavam inclusos nos dados treino. Essa classificação informa com certa probabilidade de qual classe a instância pertence. Tendo-se uma série de instâncias classificadas a estratégia é de seguir as decisões entrando no mercado quando essa diz que o preço do ativo vai subir, e ficar fora do mercado quando o preço do ativo vai descer. O desempenho dessas estratégias é demonstrado em gráficos de carteiras fictícias de investimento, onde não se levam em consideração preços de transações e nenhum outro tipo de taxa.

2 Objetivos

1. Manipular dados históricos de preços de ações, visando identificar parâmetros e padrões que atuam como informações decisivas para propor estratégias de investimentos.

2. Criar regras de decisões com algoritmos classificadores e validar os resultados criando carteiras fictícias mostrando o desempenho da estratégia perante o desempenho do ativo e índice.

3. Analisar as regas criadas e inferir quanto tempo de investimento é necessário para dobrar o valor inicial investido.

3 Métodos

3.1 Organização da Base de Dados

Os dados foram obtidos do site do Yahoo (http://br.finance.yahoo.com/), a partir dessa, foi feito o download dos preços dos valores diários das ações da PETROBRAS-PN (PETR4.SA) e dados do índice IBOVESPA. As informações contidas no Banco de dados são: data, preço de abertura, preço de fechamento, preço de máximo, preço de mínimo e volume negociado. Os dados fazem referência do início do ano de 2000, até o mês de agosto de 2010.

3.1.1 Manipulação dos dados

Para o melhor intendimento pelos programas utilizados, os dados foram trabalhados e modificados em cinco etapas. Cada etapa foi feita tanto para o índice IBOVESPA quanto para as ações da Petrobras, e estão descritas a seguir:

1. Organização dos dados de interesse;
2. Enumeração dos dados e processamento;
3. Determinação dos atributos e organização dos dados compondo as instâncias;
4. Classificação das instâncias;
5. Construindo os dados de treino;

Organização dos dados de interesse

Filtrando os dados de interesse da tabela, ou seja, de todos os dados obtido, a coluna que representa o volume de transação foi excluída ficando assim somente com os dados restantes.

Enumeração dos dados e processamento

Os dias em que a bolsa de valores não operou, foram preenchidos com os valores do último dia anterior mais próximo(que contêm informação). Esse processamento foi feito afim de manter uma memória dos dados nos dias que a bolsa de valores não abriu.

Os dados estão dispostos em N valores diários sendo informado por data qual seu dia de ocorrência. Para a enumeração substituiu-se a data enumerando a quantidade de dados existentes. Tomando-se por data de início o dia 1, e a última por N.

Determinação dos atributos e organização dos dados compondo as instâncias

Os dados foram organizados em atributos com finalidade de serem analisados pelos classificadores

A evolução temporal dos índices é não estacionário, assim sendo o valor bruto dos índices pouco informativo

Cada atributo é formado da seguinte forma:

$$A_{t\tau} = \frac{S_{t-\tau}}{S_{t-\tau-1}} \tag{3.1}$$

Onde $A_{t\tau}$ é o mesmo atributo; $S(t)$ é o preço das ações no dia t ,(t = N , ..., 2) e τ =(1,..., 5)

Os atributos que aparecem nas árvores de decisão fora rotulados a fim de identificar qual informação esse se refere. Cada nome informa se o atributo faz parte do índice ou da ação, da razão referente entre um dia $S_{t-\tau}$ e dia $S_{t-\tau-1}$, e se refera a preço de abertura, fechamento, máxima ou mínima.

A tabela com todos os dados do IBOVESPA e a tabela com os dados da Petrobras são composta por colunas que representam os atributos, e cada linha representa as Instâncias.

As instâncias são classificadas da seguinte maneira:

- Se Fechamento5I for maior que 1, classifico como “UP”,se não classifico como “Down”.

Essa classificação gera uma nova coluna onde é titulada como *Target*. Essa coluna é a variável dependente e é feita tanto para o IBOVESPA, quanto para a PETR4.

Construindo os dados de treino;

Para o treino dos dados, todos os atributos, tanto a Petrobras quanto do índice IBOVESPA, passam a fazer parte de um novo conjunto de dados, sendo o atributo target sempre a última coluna.

Quando treina-se os dados visando a estratégia para investir na PETR4, a coluna "target"do IBOVESPA é retirada. Assim como o treino para o IBOVESPA retira-se o target da Petrobras. Lembrando-se que para o treino, os atributos referentes a quinta razão (dia6/dia5), são retirados pois, a informação que a mineração de dados busca identificar é exatamente a informação do que aconteceu no mercado no dia 6.

3.1.2 Buscando padrões

Para a busca dos padrões e classificação dos desempenhos dos classificadores os dados foram processados por método de aprendizagem de máquina do software WEKA.

3.1.3 WEKA

Weka é uma coleção de algorítimos de aprendizagem de máquina provenientes de diferentes abordagens com finalidades de exploração de dados e obtenção de padrões (WITTEN; FRANK, 2000). O algoritmo pode ser diretamente aplicado a base de dados e contém ferramentas para preprocessamento de base de dados, classificadores, regressão, *clustering*, associações de regra, e visualização de dados. Esse software permite a máquina "aprender"indutivamente ou dedutivamente. Entre seus principais classificadores encontra-se aqueles geradores de árvores de decisão que é a representação de uma regra gerada em forma de árvore, ajudando a identificar melhores estratégias para se alcançar um objetivo.

3.1.4 Algoritmo J4.8

O algoritmo J4.8 é baseado no algorítimo C4.5 que é usado para gerar árvores de decisão. Esse utiliza método de entropia de informação para induzir árvores de decisão a posterior classificação.

O algorítimo cria árvores de decisão de uma base de dados de treino usando o conceito de entropia de informação. Sendo os dados de treino um conjunto $S = s_1, s_2,...$ de exemplos de dados já classificados. Cada amostra $s_i = x_1, x_2...$ é um vetor onde x_1,x_2,... representa atributos da amostra. O dados de treino são complementado com um vetor

$C = c_1$, c_2... onde c_1, c_2,...representa a classe ao qual cada amostra pertence.

Em cada nódulo da árvore, o classificador escolhe um atributo do dados que melhor divide o conjunto em em subconjuntos melhorando uma classe ou outra. Esse critério é o ganho de informação normalizada (diferença na entropia) que resulta da escolha um atributo para dividir os dados. O atributo com maior ganho de informação é escolhido para fazer a decisão (Zheng, Z.;Ting, K. 1998).

3.1.5 Valores de desempenho de classificadores

Os algorítimos de mineração de dados retornam valores que representam o desempenho das classificações naqueles dados. Para um resultado robusto, devem visar o equilíbrio das medidas de desempenho. As de maiores representatividade são:

- **Precisão**: dada pela soma dos verdadeiros positivos obtidos para todas as classes dividida pela soma de todos os verdadeiros positivos e falsos positivos;
- **Recall**: razão entre o número de verdadeiros positivos de determinada classe e número total de exemplos daquela classe;
- **ASC – Área sob a curva ROC (*"Receiver operating characteristic"*)**: A curva ROC plota a fração de verdadeiros positivos pela fração de falsos positivos, sendo que área abaixo dessa curva é numericamente igual a probabilidade de uma determinada instância ser corretamente classificada.

3.1.6 Matriz de Confusão

Formalmente a matriz de confusão é uma matriz $[c_{ij}]$, com $i = 1...N$, $j = 1...N$, onde N é o números de classe. Os elementos c_{ij} é o número de objetos da classe i que foi classificado com a classe j. É claro que se $i = j$ o objeto da classe i foi classificado como sendo da classe i (Susmaga R. 2007).

Matriz de confusão é muito usada para avaliar classificadores, assim essa proporciona um resultado bem rápido a respeito da performasse exibindo em forma de matriz a distribuição das instâncias classificadas corretamente e classificada erroneamente (Susmaga R. 2007).

As colunas representam as instâncias que foram classificadas de acordo com uma das classes. As linhas representam quantas instâncias pertence a cada classe. A soma da

diagonal principal representa o total de classes que foram classificadas corretamente em suas respectivas classes, enquanto a soma da da diagonal secundaria representa o total de instâncias classificadas erroneamente.

3.1.7 Validação Cruzada

Como costuma-se trabalhar com amostras, é interessante dispormos de ferramentas que possam estatisticamente validar os valores obtidos e os modelos gerados. O método de validação cruzada por k vezes consiste no particionamento aleatório da amostra em k subamostras (geralmente, $k = 10$). Uma única subamostra é separada para o teste do modelo, enquanto as restantes são utilizadas para o treino. O processo é repetido até que as k subamostras tenham sido utilizadas para teste. Os valores de precisão, recall e ASC que o algoritmo retorna é a média obtida para os k testes (PICARD; COOK, 1984).

(Validação Cruzada) **Divisão dos dados para gerar as carteiras.**

Sendo o número total de instância NI_i, para o treinamento da primeira carteira utilizou-se NI_{i-50} dados. As últimas 50 informações são usadas para serem treinadas pelo modelo. Para a segunda carteira os 50 dados que foram testados na primeira são desconsiderados, fazendo assim que os dados de treino sejam N_{i-100} e os de teste sejam de N_{i-100} à N_{i-50}. Assim consecutivamente até o treino de N_{i-450} e informações para teste de N_{i-450} a N_{i-400} que constitui a última carteira.

A carteira é composta da seguinte forma.

1. O valor inicial V_0 $(t = 0)$ é 1;

2. Se a regra me diz que o mercado vai subir no dia seguinte, esse V_0 é multiplicado pelo ganho que esse dia analisado teve realmente no Índice ou na ação.

3. Se minha classificação for “Down”, esse valor é multiplicado por 1, simulando ficar fora do mercado.

4. Esse valor multiplicado, é mantido na memória e caso o próximo dia analisado ser novamente “Up”, esse é multiplicado mais uma vez pelo valor real do índice. Fazendo assim com que a carteira mostre um gráfico do ganho acumulado durante 50 dias.

3.1.8 Comparação dos classificadores

Uma quantidade de 450 dados foram classificados por diferentes algorítimos para comparar seu desempenho afim de melhorar o ganho obtido com a estratégia. O processo de classificação é o mesmo descrito e a carteira é formada da mesma maneira que para as anteriores. Esses gráficos são plotados e analisados por meio de ajuste linear permitindo inferir quanto tempo de investimento é preciso para retornar o valor inicial.

4 Resultados e Discussão

Após a mineração e organização dos dados em listas que podem ser processados pelo software Weka, um modelo gerado pelos dados treino foi usado para classificar os dados testes. Seguindo a regra criada, simulou-se uma carteira de investimento para validar os resultados tanto para o Ibovespa quanto para a Petrobras.

O algoritmo J4.8 gera uma árvore de decisão para mostrar qual regra utilizada foi seguida para ser feita a melhor classificação dos dados. Os atributos que são ditos ter maior ganho de informação podem ser identificados com a visualização dos histogramas das suas classes, como na figura 1, onde na figura 1-a e 1-b representam um atributo com poucas informações distintas e um atributo com muitas informações distintas respectivamente. A barra do histograma que contém o valor número 1 é mais expressivas que as outras pelo motivo da grande quantidade de dados que foram preenchidos com os valores anteriores a fim de manter a memória em dias que o índice não foi cotado.

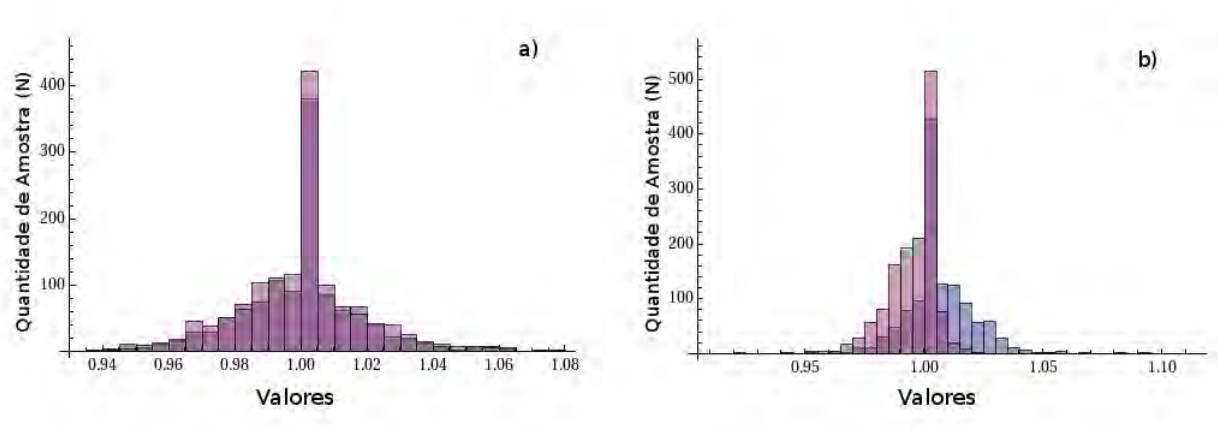

Figura 1: Histogramas representando atributos com ganho de informação: (a) pouco ganho de informação; (b) grande ganho de informação.

As árvores geradas pelos algoritmos, apresentam características diferentes quando treinadas com dados em que as variáveis dependentes dizem a respeito do movimento da

PETR4 e do IBOVESPA. Isso se deve não somente, mas também ao fato de que o índice PETR4 é somente um ativo enquanto o índice Ibovespa corresponde a uma média do desempenho de todos os ativos da bolsa.

4.1 IBOVESPA

Segundo a árvore de decisão gerada pelos dados classificados com os eventos ocorrido no IBOVESPA, figura 2, foram utilizadas 4 informações "atributos"para gerar a regra de decisão: O Máxima4I , Mínima4I, Abertura2I e o Fechamento4I.

Suponha a classificação seguindo a regra gerada, e tendo que ser decidido qual o comportamento a ser tomado pra a movimentação no mercado no $t + 5$ dia, o atributo mais relevante a ser observado é o Maxima4I. Este se encontra no topo da árvore e contém maiores informações distintas de cada classe.

Um exemplo de como seguir a árvore é: Se o valor Máximo do índice IBOVESPA do dia 4 for menor 1.008347, e o Valor Mínimo do Índice IBOVESPA do dia 4 for maior que 0.990142; Nos caso avaliados, 519 casos foram classificados como "Down", o preço de fechamento do sexto dia é menor que o preço de fechamento do quinto dia, e 59 foram classificadas de forma errada.

O que esta regra de classificação diz é que leva-se em consideração informações tanto de preço de abertura quanto de fechamento quanto máxima e mínima do índice para uma análise a fim de saber qual a direção do mercado no sexto dia. Percebe-se que a razão das informações contidas em cada folha, que se encontram mais próximas do ápice, são maiores das que se encontram inferiormente na árvore. Isso diz que quanto mais alta a folha mais valiosa é a informação e a classificação.

Nesta regra, ambas as classes foram compreendidas pelo classificador de uma mesma forma já que a quantidade de "Up"é igual a quantidade de "Down". A informação do preço de máxima do 4^0 dia do Ibovespa é o que tem mais importância e é o que deve ser observado primeiramente segundo a regra. Os preços de fechamento compõem as folhas finais da árvore, sendo a última informação a ser levada em consideração para a tomada de decisão.

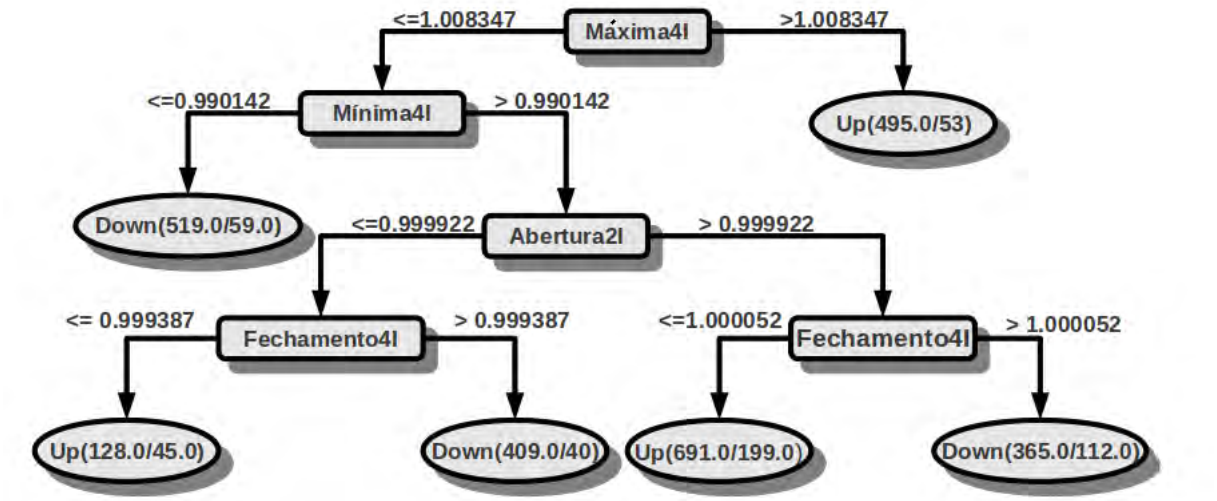

Figura 2: Árvore de decisão gerada pelo algoritmo J4.8 para os dados de treino da Ibovespa.

A matriz de confusão gerada para o conjunto de treino do Ibovespa corresponde à tabela 1. A diagonal principal contêm os números referentes a classificações corretas de cada classe, sua soma é referente a quantidade de informação classificada como correta dentro de todos os dados. Assim, percebe-se que o desempenho do algoritmo é considerável e que esse conseguiu distinguir com certa clareza uma grande quantidade de dados. Nesse treino temos uma precisão maior para a classificação da classe "Down", pois a razão entre classes corretamente classificadas por erroneamente classificadas é maior do que para a classe "Up". Isso pode-se basear na idéia de que padrões que se encontram em tendência de baixa são encontrados mais facilmente pelo classificador.

Tabela 1: Tabela com os resultados obtidos durante o treinamento do modelo.

Down	Up
1124	255
319	909

Os resultados do desempenho do classificador são mostrados em formas de tabela, onde estão os valores mais importantes como: *Recall*, *Precision* e Área sob a curva ROC. Os valore de *Recall* e *Precision* dos modelos gerados para classificação do movimento do mercado (0.779 e 0.815 respectivamente), indicam uma boa classificação contendo alguns ruídos que dizem ser ambigüidade nas classificações. Isso se deve ao fato de que

o movimento do ativo não depende estritamente das características de seus valores nos últimos dias, e se caso depender, o mesmo movimento pode indicar tanto alta como baixa. O Recall diferente para as duas classes significa mais uma vez que as instâncias classificadas como Down são classificadas com mais precisão do que as classificadas para a classe, significando exclusividade dos padrões que compõe a classe Down.

Tabela 2: Valores do desempenho do classificador J4.8 para os dados do Ibovespa.

Precision	*Recall*	*ROC Area*	*Class*
0.779	0.815	0.817	Down
0.781	0.74	0.817	Up

4.2 Petrobras

Para o caso da árvore gerada pro índice PETR4, figura 3 , observa-se somente 3 atributos escolhidos para compor a regra de decisão: Abertura4P, Abertura2P, Fechamento4P.

O que se percebe pela árvore é que a classe "Up"foi mais identificada influenciando mais da regra de decisão. Para esse modelo é mais minuciosa a classificação do movimento de alta do que movimento de baixa. A quantidade real de altas do verdadeiro movimento do índice pode influenciar e mudar os atributos escolhidos pela árvore, pois quanto maior a quantidade de padrões que a máquina possa aprender, mais preciso é o modelo.

A matriz de confusão para o PETR4 é vista na tabela 3, e percebe-se que a precisão da classe "Down"é maior que a precisão da classe "Up", vendo que acertou-se mais o movimento de baixa do que o movimento de alta. A soma da diagonal secundária representa 36% da soma da diagonal primaria, mostrando que a classificação feita corretamente é bem eficaz.

Tabela 3: Tabela representando a Matriz de confusão gerada pelo classificador J4.8 para a Petrobras.

Down	Up
1009	323
355	871

Os valores de Recall e Precision dizem que o classificador acerta uma quantidade bem próximas de classes classificadas dentro das realmente pertencentes às classes, tanto para

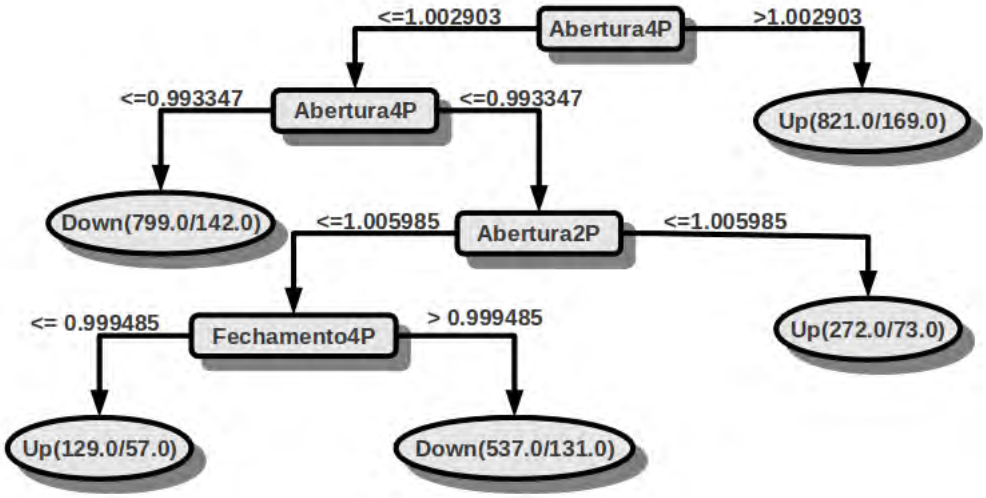

Figura 3: Árvore de decisão gerada pelo algoritmo J4.8 para os dados de treino da Petrobras.

"Up", quanto para "Down". A curva ROC, é considerada ótima a partir do valor 0.8. Para o índice PETR4 o valor de 0.777 está dentro de uma faixa considerável para validar o método.

Tabela 4: Valores do desempenho do classificador J4.8 para os dados da Petrobras.

Precision	Recall	ROC Area	Class
0.74	0.758	0.777	Dow
0.729	0.71	0.777	Up

4.3 Carteiras

Utilizando-se da regra gerada, o gráfico da carteira do IBOVESPA com o ganho acumulado é mostrado na figura 4, junto com os gráficos utilizado para a comparação do desempenho em relação a: Desempenho real do ativo, desempenho de um classificador perfeito (nunca erra), desempenho de um classificador imperfeito (sempre erra), desempenho do classificador j48 imperfeito (todos os acertos são considerados erros, e todos os erros são considerado acertos).

Com uma taxa de acerto por volta dos 76%, o gráfico gerado considerando a classificação do J4.8 tem um comportamento na maioria das vezes parecido com o compor-

tamento do gráfico gerado pelo classificador perfeito. Espera-se que sempre a linha do classificador fique acima das linhas do classificador imperfeito e do classificador J48 imperfeito e abaixo do classificador perfeito.

Percebe-se em algumas partes do gráfico onde é que o classificador se diferencia do real comportamento do ativo. Por volta do dia 26 até o dia 32 da (figura 4), o ativo acumulou uma queda e essa, foi compreendida pelo algoritmo classificando-a como baixa do mercado, assim a decisão de "ficar fora do mercado"foi tomada fazendo com que a perda fosse nula. Do dia 32 ao dia 40 aproximadamente, o gráfico mostra que houve uma alta. Esse mais uma vez foi entendido pelo classificador, onde classificou o movimento de alta do ativo. Assim a decisão tomada foi de compra, acumulando um ganho notável. No final do período o gráfico mostra o alto desempenho que a regra gerada pelo J4.8 criou. O método dá assim, uma credibilidade para esse tipo de análise afirmando que a estratégia é válida e gera bons resultados.

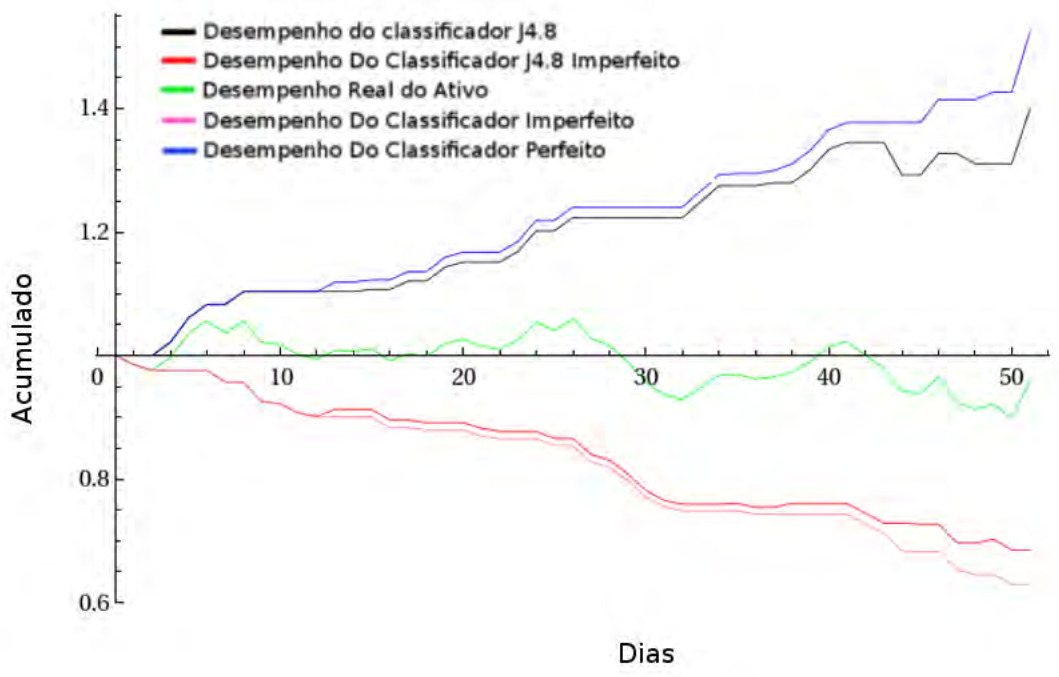

Figura 4: Carteira simulando o investimento seguindo a regra gerada pelo classificador para os dados do Ibovespa.

Para o gráfico produzido utilizando a regra com a base de dados referente ao PETR4, obteve-se o gráfico da figura 5, junto com os mesmo gráficos para comparação citado no gráfico da figura 4.

Neste, percebe-se uma distância maior da linha final do classificador que segue o

classificador perfeito. Isso se deve a classificações incorretas feita quando o mercado assumiu uma postura de baixa e foi interpretada como uma postura de alta, assim a ordem dada foi de entrar no mercado passando a acumulando um ganho negativo.

A alta ocorrida entre o dia 22 e 27 foi compreendida pelo classificar que tomou a forma de uma tendência de alta nesses dias juntamente com a tendência de alta que o ativo teve realmente. Na proximidade do dia 4 é notável que o algoritmo interpretou uma alta brusca no mercado, acumulando ganho, seguida da interpretação da baixa brusca minimizando a perda.

O ganho total neste período terminou mais uma vez maior que o ganho do real movimento do ativo. Isso é esperado, pois com um acerto de cerca de 72% de instâncias corretamente classificadas, espera-se que a regra associada elimine as baixas mantendo-se fora do mercado e maximize o ganho "entrando no mercado".

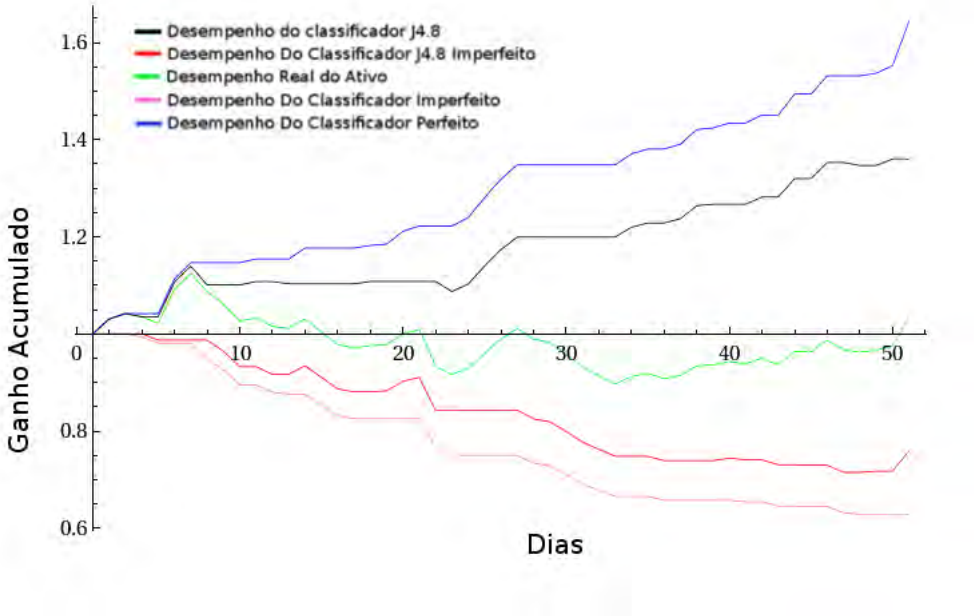

Figura 5: Carteira simulando o investimento seguindo a regra gerada pelo classificador para os dados do Ibovespa.

4.4 Desempenho dos classificadores

Vários algorítimos foram usados para treinar a mesma base de dados, e todos esses geraram uma regra de decisão diferente. Para determinação do melhor classificador, a comparação das carteiras foi gerada e os resultados analisados. Dentre os classificadores usados estão: J4.8, *LMT*, *Logistic*, *ADTree*, *FT*, J4.8 com *Bagging*.

O gráfico obtido ao longo da análise de 450 dias é demonstrado na figura 6. Apesar do classificador J4.8 apresentar resultados inferiores aos outros, esse informa a regra de decisão de forma mais clara e fácil de identificar os padrões. Os outros classificadores por terem seu funcionamento diferente, apresentaram um desempenho melhor dizendo que: pode-se melhorar a análise e as identificações dos padrões que estavam na base de dados.

O gráfico demonstra que para uma futura análise dos dados o algoritmo *Logistic* gera um ganho maior, aumentando assim, as chances de acertar o movimento a ser predito.

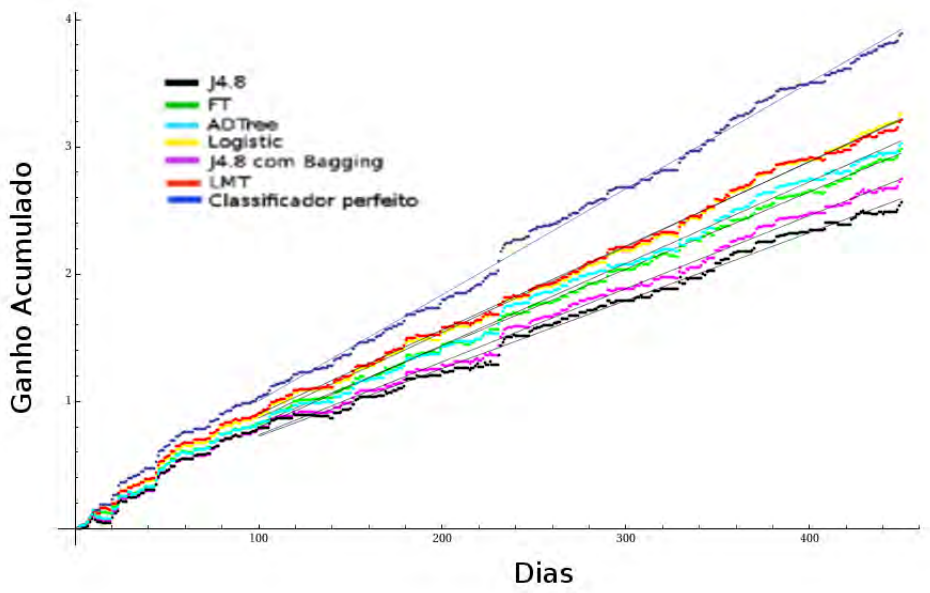

Figura 6: Gráfico que compara o desempenho dos diferentes tipos de classificadores usados para gerar a regra de decisão.

O desempenho dos classificadores são demonstrados por ordem de relevância na tabela 5.

Tabela 5: Desempenho dos diferentes classificadores usados na figura 11.

Classificador	Precision	Recall	ROC
LMT	0.873	0.87	0.902
Logistic	0.82	0.864	0.893
ADTree	0.804	0.869	0.855
FT	0.807	0.8227	0.823
J4.8 com Bagging	0.789	0.859	0.866
J4.8	0.793	0.849	0.822

Foi estabelecido que nos primeiros 100 dias a curva se distancia do comportamento linear e por isso o modelo do ajuste foi feito somente para os últimos 350 dias.

O ganho acumulado tem um comportamento exponencial, assim pra poder ajustar a curva passamos os valores do eixo y para $Log[y]$. O ajuste linear foi feito com o modelo $y = A + bx$, onde o b encontrado é a taxa de crescimento. Segundo os ajustes lineares encontrados no desempenho dos classificadores, é possível inferir quanto tempo é necessário para alcançar o dobro do valor inicial e qual classificador retorna lucro mais rapidamente. Considerando a equação:

$$In = In_0 * e^{bt} \tag{4.1}$$

onde In_0 é o valor inicial de investimento, In é o valor do investimento depois de t dias, b é a taxa de cescimento encontrada pelo ajuste linear e t é o tempo de investimento (dias), estimou-se para cada classificador o tempo t necessário para conseguir o dobro do valor inicial ($2In_0$) Substituindo In por $2In_0$ chegmos a relação

$$t = \frac{ln2}{b} \tag{4.2}$$

A taxa de crecimento de cada classificador e em quanto dias de investimento me retorna o dobro do investimento inicial estão na tabela 6.

Tabela 6: Desempenho dos diferentes classificadores.

Classificador	Taxa de crescimento	Dias para dobrar o valor inicial de investimento
Clasificador perfeito	0.00833	83
Logistic	0.006722	104
LMT	0.00666	106
ADTree	0.00643	108
FT	0.00606	115
J4.8 com Bagging	0.00933	121
J4.8	0.00933	130

Com a taxa de crescimento encontrada pelo modelo linear comprovou-se que os classificadores que tem melhor desempenho são os que retornam um ganho mais rapidamente. Os modelos logísticos são mais precisos percebendo também que o *Bagging* melhora a classificação do algoritmo J4.8.

5 Conclusão

Visando validar um método dinâmico e econômico que retorne lucro no investimento, propomos um método computacional simulando estratégias de investimentos baseadas em termos e padrões descobertos em base de dados.

Comprovam-se as teorias de que o movimento do mercado sempre se repete. Caso isso não acontecesse, o classificador não encontraria padrões suficientes para criar regras e seu desempenho seria de um classificador aleatório. Possivelmente uma análise de uma janela de tempo diferente possa descobrir outros padrões e termos ajudando a compreender o mercado financeiro. Assim como outras análises feitas com diferentes classificadores também pode ajudar na tarefa de identificação de regras e tendências.

Pela visualização das árvores percebe-se que:

- Ambas apresentam uma melhor classificação para a classe Down. Isso significa um melhor entendimento pela máquina dos padrões quando o mercado está em tendência de queda. Isso significa que o comportamento dos investidores em ocasiões que o mercado está em baixa é menos diversificado do que quando esta em alta.

- O classificador mostrou-se bem eficaz e essa característica é comprovada quando obteve-se um ganho maior do que o ganho do desempenho real do Ativo para todas as carteiras. Os valores de *Recall* e *Precision* fortalecem e comprovam um método fácil e rápido de gerar boas decisões no mercado financeiro.

Lembrando-se que outras variáveis são muito importantes em um investimento, os resultados obtidos das carteiras com princípio de estudos dos padrões contidos nas base de dados, mostra-se satisfatória para o tipo de análise feita. Teoricamente, a estratégia me retorna um desempenho sempre melhor que o real desempenho do ativo, mostrando-se bem eficaz validando a técnica. Os modelos lineares conseguem me dizer qual a taxa de crescimento do investimento podendo assim estimar o tempo que é preciso para obter um lucro do dobro do valor de investimento inicial, possibilitando um investidor a se

organizar e estimar quanto tempo leva para alcançar um determinado retorno. O modelo linear comprova o desempenho do melhor classificador, mostrando assim que é possível seguir esse tipo de análise e investir segundo essas regras e no final, obter lucro.

Referências

BATTITI, R. Using Mutual Information for Selecting Features in Supervised Neural Net Learning, IEEE Trans. Neural Networks, vol. 5, no 4, pp. 537-550, Jul. 1994.

MANTEGNA, R.; STANLEY, H; AN INTRODUCTION TO ECONOPHYSICS Correlation and Complexity in Finance, Cambridge Univ. Press.

WITTEN, I.; FRANK, E. Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. 2a . ed. São Francisco: Morgan Kaufmann Publishers, 2000.

PICARD, R.; COOK, R. Cross-validation of regression-models. Journal of the American Statistical Association, v. 79, n. 387, p. 575?583, 1984.

SUSMAGA.R; Advances in software computing, intelligent information processing and Web Mining

Alex, F(2002). Data Mining and knowledge discovery with evolutionary algorithm. Natural Computing Series. Springer, (2002).

Zemke, S; Data Mining for Prediction Financial Series Case. 2003. Doctoral Thesis - The Royal Institute of Technology, Department of Computer and Systems Science,Sweden, December 2003.

Zheng, Z., Webb, G., & Ting, K. (1998). Integrating boositng and stochastic attribute selection committees for further improving the performance of decission tree learning (Technical Report). School of Computing and Mathematics, Deakin University, Geelong, Australia.

Swingler, K. (1994). Financial prediction, some pointers, pitfalls and com- mon errors (Technical Report). Centre for Cognitive and Computational Neuroscience, Stirling Univ., UK.

6 Apêndice

6.1 Apêndice A: Definições em Mineração de Dados

Alguns termos são utilizados com frequência durante a mineração de dados, e faz-se necessário defini-los corretamente:

- **Instância**: Objeto a ser classificado, independente do conceito a ser aprendido;
- **Atributos**: Características que descrevem determinado conjunto de instâncias. Quando várias instâncias apresentam determinado atributo com mesmo valor, dizemos que tais instâncias pertencem à mesma *Classe* para aquele atributo;
- **Dado**: Sequência de símbolos quantificados ou quantificáveis, para determinado atributo. Determina a *classificação* da respectiva instância;
- **Treino**: Etapa na qual o algoritmo busca as regras de associação entre os dados disponibilizados;
- **Modelo**: Conjunto de regras que buscam determinar corretamente a classificação de determinada instância;
- **Verdadeiros Positivos (*Vp*)**: Instâncias corretamente classificadas como pertencentes a determinada classe;
- **Verdadeiros Negativos (*Vn*)**:Instâncias corretamente classificadas como não pertencentes a determinada classe;
- **Falsos Positivos (*Fp*)**: Instâncias erroneamente classificadas como pertencentes a determinada classe;
- **Falsos Negativos (*Fn*)**: Instâncias erroneamente classificadas como não pertencentes a determinada classe.