Predição de mínimos e máximos locais para investimento em bolsa de valores utilizando aprendizado de máquina

Daiane Sampaio Santos

A Deus,

À minha mãe Francisca,
Ao meu pai Luiz (em memória)
Ao meu irmão Maxwell.

Agradecimentos

Antes de mais nada e acima de qualquer coisa, agradeço a Deus, que na Sua infinita bondade não desistiu de mim, me amou e me ensinou o Seu amor. Me concedeu o dom da vida e a graça de vivê-la com pessoas tão maravilhosas ao meu lado.

Agradeço à minha família, à cada um daqueles que, de uma maneira ou outra, participaram de perto dessa jornada. À minha mãe Francisca, o meu profundo agradecimento. Pela força, coragem e determinação. Ela que mesmo em meio à todas as dificuldades, me proporcionou a chance de ir além. Despertou e desperta em mim o gosto pelo estudo, pelo conhecimento e pela vida. Ao meu irmão Maxwell, pela inspiração, calma, dedicação e paciência. Por ter contribuído com a minha escolha pela computação, e mais ainda pelo mestrado. Ao meu pai Luiz, que descansa em paz com Deus, por ter me deixado o legado da arte de sonhar.

Agradeço ao meu orientador Prof. Dr. Edson Takashi Matsubara, por tudo o que ele cuidadosamente se empenhou para que eu aprendesse. Pela sua dedicação profissional e prazer pela pesquisa, que me motivaram a querer aprender sempre mais. Pela bondade e humildade do seu coração. Pelos conselhos, pela amizade, pelo exemplo de vida, o meu muito obrigado.

Agradeço aos colegas do LIA, Zezeu, Dadau, Capi, Guiguis, Adolto, Manuel, Mari, Carlos, Dani, Eduardo's, Lucas's, Juninho, Anderson, Yumi, Bruno, Yuri, Greyce, Paulo, Marcos, Rodrigo, Pedro, pelo bom convívio, pelas experiências divididas e pelo conhecimento compartilhado, pelas confraternizações e churrascos, pela batata com cheddar e bacon, pelas partidas de Catan e Agrícola, pelo Butteryaki, pela padoca (Ah! a padoca), pelos aniversários surpresas (e os chapeizinhos amarelos), pelo hambúrguer com batata frita, e por tantos momentos juntos. Agradeço ao Hudson, Koiti e Eduardo Max pelo grande auxílio neste projeto.

Agradeço aos meus amigos pelo apoio e carinho durante esse tempo de mestrado. Pelas vezes que me apoiaram, comemoraram comigo, e pelas vezes

que me consolaram nos momentos de desespero, sem vocês teria sido difícil. Agradeço à Suellen, uma amiga de longa data, que com a sua família, que também considero minha, tornam a minha vida mais gostosa (o que seria de mim sem as jantas de quarta e almoços de domingo? e sem os chocolates, mousses e todo amor que eu sempre senti de vocês?). Agradeço à Cah, Mari e Jota, pela boa compahia, pessoal e virtual, e por fazeram dos nossos encontros verdadeiros retiros de risadas. Agradeço à Liga, ao Júlio, César e a família de votermincês, que me conheceram no meio desse caminho e se tornaram fundamentais, tantas e tantas vezes. Agradeço ao Diego, aquele amigo remoto que suportou os meus surtos de vez em quando. Agradeço à Ester, que do jeito dela sempre buscou levantar o meu astral. Agradeço ao Gaúcho, meu fiel escudeiro, aquele amigo pra chamar de meu. Agradeço a todos os meus amigos, Borges, Michela, Paula, Luiza, e todos aqueles que eu possa ter esquecido de citar aqui. O meu muito obrigado.

Agradeço também aos colegas do mestrado, que dividiram as angústias, dúvidas, medos e prazos. Ao casal nota dez Vanessa e Fabrício, que proporcionou os melhores churrascos da turma. Ao Paulo e Kleber, pelo trabalho de SO que nos uniu. Ao Marcel, Joelmo, Virmerson, Roní, Pietro, Alex, Simone, Tatiane, Irani e tantos outros, por sonharem os mesmo sonhos e tornarem as disciplinas mais divertidas e compensadoras.

Agradeço ao pessoal da pós-graduação da FACOM, aos professores, secretários, coordenadores e todos aqueles que fazem parte dessa incrível instituição. Por fim, gostaria de agradecer à CAPES pela minha bolsa de mestrado e à FACOM-UFMS pela estrutura e apoio disponibilizados durante minha formação.

Resumo

A análise de tendências de preço no mercado de financeiro requer elevada atenção do analista de mercado quanto às variáveis que podem influenciar o preço das ações. As corretoras que atuam na bolsa de valores investem recursos em análises financeiras, para em troca obterem recomendações de compra e venda de ações. O desafio dos analistas consiste em sinalizar a compra e venda das ações, de modo a maximizar os lucros. Nesse sentido, a predição de ações tem sido foco de constantes estudos. Muitos argumentam da impossibilidade de criar modelos capazes de predizer o comportamento de um ambiente tão instável e com tantas variáveis. Entretanto, algoritmos de Aprendizado de Máquina (AM) são apropriados para situações com diversas variáveis e padrões a serem descobertos. Para tanto, as informações financeiras dispostas em séries temporais são transformadas em tabelas atributo valor, para que se adequem ao formato de entrada dos algoritmos de AM. Quanto à essa transformação, a literatura têm sugerido a utilização de indicadores econômicos para predição da tendência futura de preço absoluta. Entretanto, acredita-se que uma maneira mais significativa de representar a classe do problema seja baseada em valores máximos e mínimos da série temporal. Nesse sentido, este trabalho propõe uma representação de classe denominada LMinMAx, que estima pontos de máximo e mínimo e os utiliza como atributos classe nos conjuntos de dados. Os experimentos desenvolvidos comparam a abordagem proposta com outras duas representações de classe propostas na literatura e, em termos financeiros, com carteiras recomendadas e aplicação em poupança. Os resultados são promissores e mostram que a abordagem proposta pode ser utilizada para recomendação automática de compra e venda de ações. A abordagem proposta supera as principais representações de classe com diferença significativa ($p = 0.05$) em termos de AUC e rendimento.

Sumário

Lista de Figuras

Lista de Tabelas

CAPÍTULO

1

Introdução

O *capital social* de uma empresa pode ser parcialmente colocado à venda, para o público em geral, sob a forma de *ações*. Os compradores das ações, ou *investidores*, pagam o preço cotado na bolsa de valores, e se tornam sócios, ainda que de forma minoritária, da empresa da qual a ação foi comprada. Empresas que permitem a compra e venda de suas ações são ditas de *capital aberto*. Os *corretores* cadastrados na bolsa são pessoas (físicas ou jurídicas) autorizadas a executar operações de compra e venda de ações, de acordo com as ordens de seus clientes, que são os investidores.

O investimento em ações tem sido amplamente praticado por especialistas financeiros, e até mesmo leigos no assunto, que analisam o mercado financeiro em busca de informações sobre tendência de subida ou descida das ações. Entretanto, a decisão de quando comprar (subida) e vender (descida) ainda é uma tarefa bastante difícil.

Para auxiliar os investidores, algumas corretoras mantém *carteiras recomendadas* de ações, que são sugestões periódicas de compra e venda focadas na obtenção de lucro a médio prazo. As carteiras recomendadas são fruto da análise minuciosa do mercado, e seu rendimento varia bastante, entre os anos e entre diferentes corretoras. A lista de rentabilidade anual das melhores carteiras para a BOVESPA, ilustrada na Tabela 1.1, mostra que dificilmente a mesma corretora aparece no topo, em anos consecutivos. Assim, é difícil para o investidor selecionar quais recomendações seguir.

Corretora	2011 (%)	Corretora	2012 (%)	Corretora	2013 (%)
Souza Barros	6.74	Geral	55.48	Ágora/Bradesco	15.30
HSBC	-3.41	Alpes/Wintrade	52.81	Geração Futuro	10.40
Planner	-5.38	Souza Barros	52.04	BTG Pactual	7.30
BTG Pactual	-6.60	Coinvalores	44.60	BB	2.43
WinTrade	-11.25	Fator	42.60	Pax	2.34
XP	-12.10	Rico/Octo	36.11	Rico/Octo	1.66
BB-BBI	-12.90	BI&P	35.26	Geral	0.93
Ativa	-13.60	XP	34.50	Concórdia	-0.35
Gradual	-16.30	BTG Pactual	34.40	HSBC	-2.20
Socopa	-17.53	Pax Corretora	34.08	Coinvalores	-2.22
Média 2011	-9.23	Média 2012	42.19	Média 2013	3.56

Tabela 1.1: Rendimento anual das TOP 10 carteiras recomendadas dos anos de 2011, 2012 e 2013

1.1 Motivação

No sentido de auxiliar no investimento em ações, algoritmos de Aprendizado de Máquina (AM) são utilizados para extração de conhecimento sobre os históricos de cotações. Uma das tarefas da mineração de dados financeiros é identificar padrões nos preços das ações, que podem servir como recomendações de compra e venda para os investidores. Entretanto, os históricos de cotações são naturalmente dispostos em séries temporais, que são um formato não compatível com a entrada da maioria dos algoritmos de AM. Essa restrição implica na necessidade de transformar a série temporal financeira em uma representação de dados no formato de tabela atributo valor, como ilustrado na Figura 1.1.

Uma das características desejáveis da transformação, ilustrada por uma nuvem na Figura 1.1, é a capacidade de mapear a relação temporal entre os valores da série para os atributos. Assim, mesmo tratando os dados como independentes entre si, os algoritmos de AM podem identificar possíveis padrões entre preços próximos cronologicamente. Nesse sentido, a literatura tem reportado algumas transformações que utilizam indicadores econômicos, cujos valores são resultado de cálculos de medidas como médias e diferenças de preço sobre as cotações de um período de tempo.

1.2 Objetivos

Além da representação dos atributos, a tabela atributo valor também é composta pela representação de classe, a qual associa rótulos à conjuntos de atributos de acordo com as categorias do problema de predição. Especificamente na predição de ações, as classes podem ser convertidas para operações

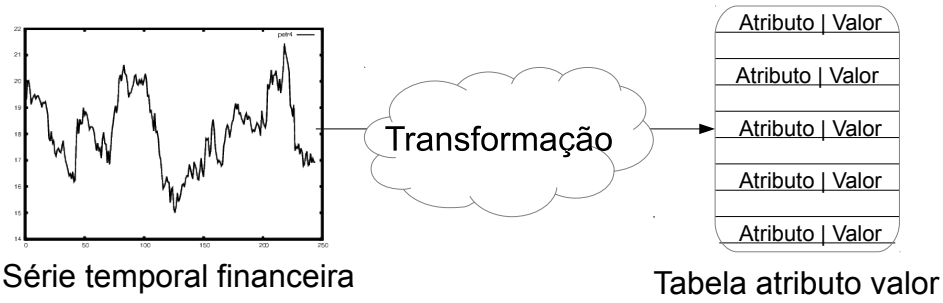

Figura 1.1: Transformação da série temporal financeira em tabela atributo valor

de compra e venda de ações.

Os trabalhos encontrados na literatura sobre predição de ações usando aprendizado de máquina têm reportado a representação de classe ou como um problema de regressão, ou como classificação de tendência de preço futura. Entretanto, dessa maneira variações de preço irrelevantes podem ser tendenciosas e, consequentemente, podem sinalizar operações de compra e venda de ações que podem não compensar devido ao custo associado.

Em geral, deseja-se que as operações sejam sinalizadas em momentos nos quais o preço é o menor possível (compra) e o maior possível (venda). Para tanto, a seleção de preços mínimos e máximos na série temporal financeira pode possibilitar um desempenho mais interessante. As ações podem ser compradas nos dias cujo preço é mínimo e vendidas nos dias de máximo, de forma a maximizar o lucro.

1.3 Contribuições

A fim de encontrar dias bons de compra e venda de ações, este trabalho propõe uma técnica para encontrar pontos de máximo e mínimo na série temporal financeira, utilizando uma janela deslizante sobre as cotações. São construídas representações de dados com exemplos diários, compostos por indicadores econômicos e cujos atributos classe são a indicação de pontos de máximo ou de mínimo. Os algoritmos de aprendizado de máquina K-vizinhos

mais próximos (KNN) e Máquina de Vetores de Suporte (SVM) são utilizados para induzir classificadores, cujas predições servem para simular operações de compra e venda. A etapa de simulação busca ser mais próxima possível do real, com taxas de corretagem e compra e venda de quantidades exatas de ações. Os resultados mostram que a abordagem proposta pode superar estratégias financeiras consolidadas, como carteiras recomendadas, com rentabilidade média anual de até 6.95%.

1.4 Organização

Para descrever a abordagem proposta e os experimentos realizados, os capítulos deste trabalho estão organizados como a seguir:

Capítulo 2: Aprendizado de máquina

Neste capítulo são introduzidos conceitos, notações e termos de aprendizado de máquina, tais como os algoritmos k-vizinhos mais próximos e máquinas de vetores de suporte, técnica de calibração e medidas de desempenho de classificadores.

Capítulo 3: Mercado financeiro

Neste capítulo são introduzidos conceitos, notações e termos de mercado financeiro, tais como os indicadores econômicos e medidas de desempenho de investimentos.

Capítulo 4: Representação de classe em mercado financeiro

Neste capítulo são apresentadas representações de classe para o problema de predição de ações no mercado financeiro usando aprendizado de máquina. Também é proposta uma nova maneira de representar as classes, utilizando mínimos e máximos locais.

Capítulo 5: Avaliação Experimental

Neste capítulo são apresentados os detalhes da avaliação experimental realizada, bem como a comparação entre os resultados esperados de cada representação de classe.

Capítulo 6: Resultados

Neste capítulo são descritos os resultados alcançados, a comparação da abordagem proposta com outras duas representações de classe descritas na literatura e comparação com estratégias conceituada do mercado financeiro.

Capítulo 7: Conclusões

Neste capítulo são apresentadas as conclusões, contribuições e trabalhos futuros.

CAPÍTULO

2

Aprendizado de máquina

Nesta seção são apresentados conceitos que fundamentam o trabalho de pesquisa, quanto aos materiais e métodos utilizados, desde a concepção dos conjuntos de dados, passando pela indução dos classificadores e chegando à avaliação de desempenho. As informações, técnicas e algoritmos apresentados a seguir são frutos de uma revisão bibliográfica sobre predição de mercado financeiro, aprendizado de máquina e outras técnicas auxiliares reportadas na literatura. São discutidos temas como informações financeiras utilizadas na mineração de dados do problema de predição de ações, técnicas que podem ser aplicadas na transformação de séries temporais financeiras em representações de dados no formato atributo valor e algoritmos de aprendizado de máquina. O capítulo está organizado da seguinte maneira: na Seção 2.1 são apresentadas algumas definições sobre representação de dados e aprendizado de máquina; na Seção 2.2 são descritos conceitos de aprendizado de máquina supervisionado, algoritmos aplicados em problemas de predição de ações e medidas de avaliação de classificadores e, por fim, na Seção 3 são descritos conceitos de mercado financeiro, tais como indicadores e índices.

2.1 Terminologia e Linguagem de Descrição de Exemplos

Nesta seção são definidos alguns termos e conceitos de aprendizado de máquina aplicados no decorrer deste trabalho, que são descritos utilizando a notação de (Edson Takashi Matsubara, 2008).

Inicialmente é apresentada a tabela no formato atributo-valor, cuja defini-

ção é de um conjunto de n_{ex} exemplos x_i com $i = 1, ..., n_{ex}$. Cada exemplo desse conjunto é composto por atributos A_j, com $j = 1, ..., n_{at}$. Sendo o valor de cada atributo, o elemento x_{ij}

Além dos atributos A_i, um atributo especial e não obrigatório de cada exemplo, denominado classe ou rótulo, é representado por y e pertence ao conjunto Y de classes discretas y_v, com $v = 1, ..., n_{cl}$.

Na Tabela 2.1 é descrita a representação de dados no formato de tabela atributo valor, com as definições de exemplo x_i, atributos A_j, valores dos atributos x_{ij}, atributos classe A_{classe} e valores dos atributos classe y_i.

	A_1	A_2	...	$A_{n_{at}}$	A_{classe}
$\mathbf{x}_1$	x_{11}	x_{12}	⋮	$x_{2n_{at}}$	y_1
$\mathbf{x}_2$	x_{21}	x_{22}	⋮	$x_{2n_{at}}$	y_2
⋮	⋮	⋮	⋱	⋮	⋮
$\mathbf{x}_{n_{ex}}$	$x_{n_{ex}1}$	$x_{n_{ex}2}$	⋮	$x_{n_{ex}n_{at}}$	$y_{n_{ex}}$

Tabela 2.1: Tabela no formato atributo valor (Edson Takashi Matsubara, 2008)

A representação de dados, por sua vez, constitui os conjuntos de dados fornecidos como entrada para algoritmos de aprendizado de máquina. Nesse sentido, os conjuntos de dados assumem dois tipos distintos:

Conjunto de treinamento: formado por exemplos rotulados apresentados ao algoritmo de aprendizado de máquina, para indução (treinamento) do classificador. Os dados que compõem o conjunto de treinamento são utilizados por algoritmos indutores, que constroem hipóteses dos dados do mundo real do qual deseja-se abstrair conhecimento.

Conjunto de teste: composto por exemplos rotulados utilizados para testar o classificador induzido. Nenhum dos exemplos que compõe o conjunto de teste deve ter sido apresentado ao algoritmo durante o processo de indução, ou seja, os conjuntos de treinamento e teste devem ser disjuntos. Assim, é possível testar qualidade do classificador, uma vez que ele não conhece os exemplos apresentados.

Como dito anteriormente, o conjunto de treinamento é utilizado para generalização do conhecimento sobre o problema de predição. Os exemplos de treinamento que o compõe podem ser rotulados, quando o atributo classe está presente, e não rotulados quando do contrário. Dessa maneira, um conceito denominado grau de supervisão categoriza o aprendizado de máquina quanto à quantidade de exemplos rotulados no conjunto de treinamento. A

Figura 2.1 mostra a variação do grau de supervisão em relação à quantidade de exemplos rotulados, categorizando o aprendizado em supervisionado, semi-supervisionado e não supervisionado.

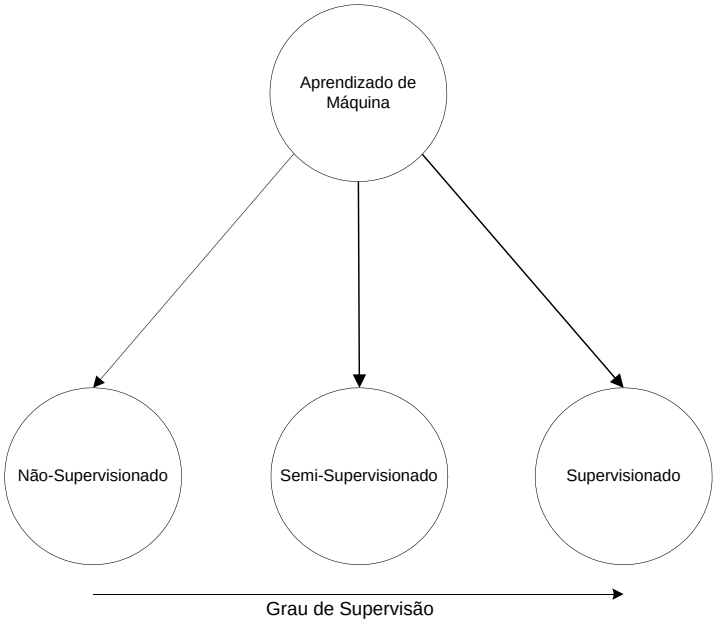

Figura 2.1: Hierarquia do aprendizado segundo o grau de supervisão dos dados (Edson Takashi Matsubara, 2008)

Independente do tipo de aprendizado no qual o problema de predição está inserido, a etapa de indução de classificadores possui alguns conceitos pertinentes, que são descritos a seguir:

Indutor algoritmo de aprendizado que utiliza um processo indutivo para gerar uma hipótese ou modelo, denominado classificador. Exemplos de algoritmos indutores são KNN, SVM, C4.5 (Quinlan, 1993) e *Backpropagation* (Rumelhart et al., 1988).

Classificador modelo ou hipótese que representa as classes de um problema e pode ser utilizado para rotular exemplos. A árvore de decisão gerada pelo C4.5 e a rede neural treinada com *Backpropagation* são exemplos de classificadores.

Super ajuste (*overfitting*) situação na qual o classificador se se superajusta aos exemplos de treinamento. Isso pode ocorrer quando a hipótese não é capaz de generalizar o conceito dos exemplos de treinamento, normalmente devido à complexidade do modelo. Consequentemente o desempenho é ruim no teste e bom no treinamento. A rede neural com uma quantidade grande de neurônios pode caracterizar um exemplo de super ajuste, e também a quantidade de vizinhos mais próximos no KNN igual a 1.

Sobre ajuste (*underfitting*) quando o classificador é muito simples e não consegue representar o conceito ocorreu. Nesse caso a classificação não é boa nem no conjunto treinamento nem no de teste. Por exemplo, quando a árvore de decisão gerada tem apenas apenas um nível (*decision stump*).

Neste trabalho os conceitos definidos acima são amplamente utilizados e o problema de predição de ações tratado é categorizado em aprendizado supervisionado. Portanto, a Seção 2.2 explica mais detalhadamente esse conceito, bem como os algoritmos de aprendizado supervisionado utilizados no trabalho e outros algoritmos encontrados na literatura, aplicados à predição de ações.

2.2 Aprendizado Supervisionado

Um problema (tarefa) pode ser resolvido utilizando aprendizado de máquina supervisionado, de acordo com o esquema apresentado na Figura 2.2. Os itens representados por elipses são tarefas realizadas por humanos, ou por máquinas (algoritmos), e os itens representados por retângulos são dados.

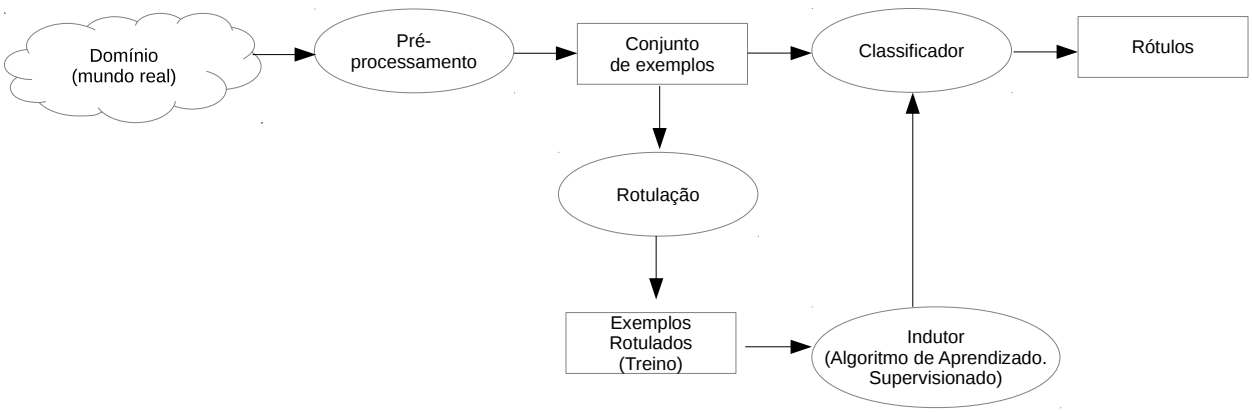

Figura 2.2: Etapas do Aprendizado Supervisionado

Para o aprendizado de máquina, um problema do mundo real precisa ser representando por uma abstração que possa ser compreendida por máquinas. Dados meteorológicos, como temperatura e pressão por exemplo, são medidos e transformados em valores numéricos, que por sua vez são tabulados e armazenados. O valor de um determinado bem, como uma casa ou um carro, pode variar de acordo com o tempo, qualidade, tamanho, entre outros parâmetros. Neste trabalho, o mapeamento do mundo real em atributos é denominado etapa de pré-processamento. Tal etapa recebe como entrada representações do mundo real e, a saída, são conjuntos de exemplos.

A saída do pré-processamento, ou seja, os exemplos compostos pelos atributos abstraídos do mundo real são rotulados de acordo com o problema de

predição para o qual foram constituídos. Para isso um especialista de domínio com conhecimento aprofundado sobre o problema, por exemplo, pode atribuir os rótulos aos exemplos. Em outros casos, o rótulo é uma característica observada nos dados que se deseja compreender, o que dispensa a rotulação manual do especialista. Por exemplo a classificação de bois com relação ao peso, problema no qual são utilizadas as classes ELITE, SUPERIOR, REGULAR e INFERIOR. Para a obtenção dos rótulos dos bois, basta saber o peso do animal e definir limiares para agrupá-los e atribuir o mesmo rótulo aos bois com peso entre dois limiares.

O algoritmo de aprendizado de máquina busca padrões (hipótese) para representar as classes, e assim os padrões podem ser submetidos ao especialista de domínio para análise e extração de padrões interessantes. Assim, a etapa de rotulação tem como entrada as instâncias e gera como saída as instâncias rotuladas. Com as instâncias rotuladas, podem ser utilizados diferentes algoritmos de aprendizado de máquina supervisionados para a indução de classificadores. Os classificadores devem ser capazes de rotular novos exemplos (instâncias), diferentes daqueles apresentados ao classificador durante a indução. A entrada dos classificadores é um conjunto de exemplos e a saída é composta por classe que categorizam cada exemplo.

Especialmente em problemas de predição de ações, a aplicação de algoritmos de aprendizado de máquina ainda tem se concentrado em dois algoritmos principais e suas modificações: Máquina de Vetores de Suporte (SVM) e Redes Neurais (NN). Em geral os trabalhos propõem diferentes configurações para os algoritmos, junção de técnicas e pré e pós-processamento.

2.2.1 Máquina de Vetores de Suporte (SVM)

O SVM pode ser considerado estado da arte em diversas tarefas de classificação e regressão. Se baseia na teoria de aprendizagem estatística de (Vapnik, 1995), que constrói um modelo que representa os exemplos como pontos em um hiper-espaço, por meio de funções *kernel* (Schölkopf, 2000). No hiper espaço são obtidos vetores de suporte que representam os limiares de decisão do algoritmo. Para realizar a classificação, um novo exemplo é mapeado para um ponto no hiper-espaço e recebe a classe de acordo com a posição do exemplo em relação ao limiar de decisão.

Dentre os trabalhos encontrados na literatura sobre predição de ações, alguns aplicam SVM na indução de classificadores para classes discretas, como (Choudhry and Garg, 2008), (Huang et al., 2005) e (Xie et al., 2006), e outros em tarefas de regressão, como (Huang and Tsai, 2009) e (Yeh et al., 2011).

(Choudhry and Garg, 2008) representa os atributos com indicadores técnicos e correlação entre ações e realiza seleção de atributos com algoritmo ge-

nético. Seus resultados chegam a 61% de taxa de acerto. Por outro lado, (Xie et al., 2006) tenta comparar o desempenho SVM com BPN e *Autoregressive Integrated Moving Average* (ARIMA), representando os atributos com séries temporais financeiras. Nesse caso, o SVM superou os outros dois métodos com 1.82% e 83.33%, de *Root Mean Square Error* (RMSE) e Estatísticas de Direção (Dsat), respectivamente.

(Huang et al., 2005) por sua vez, compara a previsibilidade dos algoritmos SVM, *Linear Discriminant Analysis*, *Quadratic Discriminant Analysis* e *Elman Backpropagation Neural Networks*. Além de propor a combinação de classificadores, a fim de unir o pontos fortes de cada algoritmo. O desempenho do SVM foi superior ao dos outros dois algoritmos, e o o modelo combinado ainda melhor, com 75% de taxa de acerto, contra 73% do SVM aplicado sozinho.

Quanto aos trabalhos de regressão, uma adaptação de SVM, denominada Regressão de Vetores de Suporte (SVR), é utilizada. (Huang and Tsai, 2009) propõe a divisão da representação de dados em fatias menores, na expectativa de diminuir a complexidade e aumentar o desempenho. O conjunto de entrada é agrupado em *clusters* com *Self-Organizing Feature Map* (SOFM) e o SVR é utilizado para induzir um modelo para cada *cluster*. Os resultados médios ficaram em 1.77 de *Mean Absolute Percentual Error* (MAPE). Em outro trabalho com SVR (Yeh et al., 2011) é proposto um método para estimação automática dos hiper-parâmetros do SVR com *Multiple Kernel*, a fim de melhorar seu desempenho. O algoritmo proposto, denominado *Multiple Kernel Support Vector Regression* (MKSVR), foi comparado com *Single Kernel Support Vector Regression* (SKSVR), *Autoregressive Integrated Moving Average* (ARIMA) e *Fuzzy Neural Networks* (FNN). Os resultados de classificação da abordagem proposta, em termos de *Root Mean Square Deviation* (RMSD), se mostraram promissores.

2.2.2 K-vizinhos Mais Próximos (KNN)

O K-vizinhos mais próximos, proposto por (Aha et al., 1991), é um algoritmo de aprendizado de máquina simples, baseado em instâncias, cujo processamento é retardado até o momento da classificação. É dito simples porque na etapa de treinamento apenas armazena os exemplos, diferentemente do SVM por exemplo, que constrói um modelo indutor com os exemplos e o utiliza para classificar novos exemplos.

Na etapa de classificação, o KNN mede a distância de similaridade entre o exemplo a ser classificado e todos os exemplos armazenados durante o treinamento. A medida de distância utilizada pode ser distância euclidiana, entretanto nada impede que outra medida seja aplicada, desde que calcule a similaridade entre os exemplos.

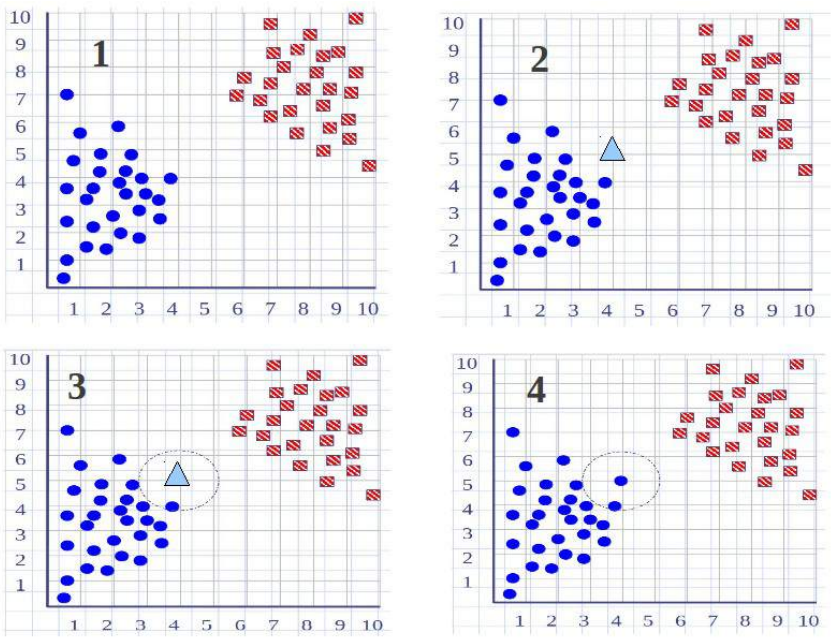

Figura 2.3: Etapas da classificação de um novo exemplo usando KNN

Basicamente, o processo de classificação de um novo exemplo x calcula a distância entre x e os todos os exemplos armazenados, seleciona a classe dos k exemplos com as menores distâncias, e a classe mais frequente dentre as k classes é atribuída a x. A Figura 2.3 mostra a ilustração dos quatro passos básicos do KNN, separados em quatro sub-figuras. O passo 1 mostra um espaço bidimensional com exemplos de treinamento armazenados e que pertencem à duas classes: "quadrado" e "bola". No passo 2, um novo exemplo representado por um triângulo é disposto no espaço, para ser classificado em uma das duas classes. A distância entre os exemplos "quadrados" e "bolas", e o novo exemplo "triângulo", permite a descoberta dos exemplos mais próximos ao "triângulo". A quantidade de vizinhos n, parâmetro do KNN, permite a formação de um raio em volta do exemplo "triângulo". Assim, os exemplos que estiverem posicionados dentro desse raio são os n mais próximos. Por fim, a classe do novo exemplo é obtida pela classe mais frequente dentre os exemplos mais próximos. Assim, o passo 4 mostra a classificação do triângulo com a classe "bola".

O Exemplo 1 a seguir mostra a aplicação do KNN a um conjunto de dados com três exemplos fictícios. Exemplo 1: considere 1-NN, ou seja, KNN com $k = 1$, e o o conjunto de dados com os exemplos da Tabela 2.2. Cada exemplo possui três atributos mais a classe. O cálculo da distância euclidiana entre os exemplos é feito com a fórmula 2.1, na qual é considerada a diferença absoluta entre os valores dos atributos. Para classificar o exemplo $Novo$ da última linha, calcula-se a distância euclidiana três vezes, entre $Novo$ e os exemplos 1, 2 e 3. Os cálculos e as distâncias finais podem ser vistos em 2.2, 2.3 e 2.4. Observa-se que o exemplo 2 tem a menor distância (1) e, por isso, é o 1-mais próximo

de $Novo$. Assim, a classe atribuída é a mais frequente do 1-mais próximo e, nesse caso, só pode ser negativa.

Número do exemplo	Atributo 1	Atributo 2	Atributo 3	Atributo classe
1	5	1	9	+
2	2	3	10	-
3	1	4	12	+
Novo	3	3	10	?

Tabela 2.2: Conjunto de dados com três exemplos rotulados e um novo exemplo a ser classificado com KNN

$$d(E_i, E_j) = \sqrt{\sum_{r=1}^{M} x_{ir} - x_{jr}{}^2} \tag{2.1}$$

$$\begin{aligned} d(E_1, E_{novo}) &= \sqrt{\sum_{r=1}^{3} (x_{1,r} - x_{novo,r})^2} \\ d(E_1, E_{novo}) &= \sqrt{(5-3)^2 + (1-3)^2 + (9-10)^2} \\ d(E_1, E_{novo}) &= 9 \end{aligned} \tag{2.2}$$

$$\begin{aligned} d(E_2, E_{novo}) &= \sqrt{\sum_{r=1}^{3} (x_{2,r} - x_{novo,r})^2} \\ d(E_2, E_{novo}) &= \sqrt{(2-3)^2 + (3-3)^2 + (10-10)^2} \\ d(E_2, E_{novo}) &= 1 \end{aligned} \tag{2.3}$$

$$\begin{aligned} d(E_3, E_{novo}) &= \sqrt{\sum_{r=1}^{3} (x_{3,r} - x_{novo,r})^2} \\ d(E_3, E_{novo}) &= \sqrt{(1-3)^2 + (4-3)^2 + (12-10)^2} \\ d(E_3, E_{novo}) &\approx 2.64 \end{aligned} \tag{2.4}$$

Embora simples, o K-vizinhos mais próximos é um algoritmo robusto a ruídos e provê bons resultados, em especial para conjuntos de dados pequenos e aplicado a séries temporais financeiras. Além da classificação em si, ele pode fornecer os exemplos mais próximos do exemplo que se deseja classificar, o que pode ser útil em aplicações que desejam não somente classificar, mas identificar características referentes aos exemplos mais próximos.

2.2.3 Outros algoritmos

Nesta seção são descritos os demais trabalhos aplicados em predição de ações, que utilizam algoritmos como árvore de decisão, indução de regras, lógica *fuzzy*, redes neurais.

Os trabalhos de (Boyacioglu and Avci, 2010), (Lai et al., 2009) e (Chang and Fan, 2008) utilizam lógica *fuzzy*. O primeiro prevê o retorno mensal do *Istanbul Stock Exchange* (ISE), com *Root Mean Squared* (RMS) de até 0.0068, utilizando *Adaptive Network-Based Fuzzy Inference System* (ANFIS). O segundo combina agrupamento em *cluster*, árvore de decisão, teoria de conjunto *fuzzy* e algoritmo genético, obtendo taxa de acerto média entre 0.7673 e 0.8688. O último compara o desempenho de Takagi Sugeno Kang (TSK) - *fuzzy-rule-based*, GA com Wang e Mendals *lgorithm for fuzzy rule generation* (GAWM), BPN e *Multiple Regression Models* (MRM). E a abordagem proposta por (Chang and Fan, 2008) obteve MAPE de 0.79 contra 1.04, 5.24 e 12.05 dos outros métodos, respectivamente.

O trabalho de (Chen et al., 2009) representa os dados com indicadores técnicos e gráficos *Candlestick*, alcançando *profit* de até 17,2%, com algoritmos de *Genetic Network Programming* (GNP) e *Sarsa Learning*

Uma revisão de artigos que utilizam redes neurais e neuro-*fuzzy* para tratar do problema de predição do mercado financeiro é realizado em (Atsalakis and Valavanis, 2009) e resume as características mais relevantes encontradas. Os trabalhos pesquisados são agrupados quanto às variáveis de entrada, metodologia de predição, modelos de comparação e medidas de performance. Os resultados mostraram que, em média, são utilizados de 4 a 10 atributos nas representações de dados, sendo mais comuns preços de abertura e fechamento, valores máximo e mínimo diários, histórico de cotações, índices e indicadores econômicos. Além disso, observou-se que uma etapa de pré-processamento dos dados é citada em quase todos os trabalhos, utilizando normalização, Principal Component Analisys (PCA), *Z-score* ou ANFIS. Por fim, a quantidade de exemplos utilizados varia entre 40 observações até 24 anos de exemplos diários.

Outros três trabalhos variam na forma como tratam os dados de entrada e aplicam redes neurais na indução dos classificadores. O primeiro - (Chang et al., 2009) - utiliza índices técnicos e *Piecewise Linear Representation* (PLR) e treina uma rede neural para detectar *trading points*, que são momentos nos quais acredita-se que a negociação da ação gera mais lucro. A taxa anual de *profit* obtida com a abordagem foi de até 35,7%. (Martinez et al., 2009) utiliza cotações passadas e indicadores econômicos Média Móvel Exponencial (MMS) e *Bollinger Bands* (BB) para treinar uma rede neural para predizer as cotações máxima e mínima diárias. Em termos de Retorno Anual (AR) e *Drawdown*,

os resultados médios ficaram em 1892,16% e 7,14%, respectivamente. Por último, (Hassan et al., 2007) propõe a fusão de *Hidden Markov Model* (HMM), Redes Neurais Artificiais (ANN) e Algoritmo Genético (GA) e a utilização de histórico de cotações. O MAPE dessa abordagem ficou entre 0,69% e 1,92%.

Por fim, o estudo de (Tsai and Hsiao, 2010) combina técnicas de seleção de atributos aplicadas à predição de ações. Os melhores resultados, em termos de precisão, foram obtidos pela intersecção de Principal Component Analysis (PCA) com algoritmo genético (79%) e multi-intersecção de PCA, algoritmo genético e árvore de decisão (78.09%).

2.2.4 Medidas para avaliação de classificadores

O desempenho dos classificadores é medido pela diferença entre a classe predita e o rótulo verdadeiro de cada exemplo do conjunto de teste. Alguns medidas normalmente aplicadas para avaliação de desempenho de classificação são acurácia, precisão, erro e revocação. Além disso, uma técnica denominada Análise *Receiver Operating Characteristic* (ROC) tem sido aplicada na avaliação dos classificadores, bem como na calibração de limiares de decisão.

Medidas de desempenho primárias

A tradução das medidas de desempenho primárias da linha inglesa para a portuguesa tem significativa diferença. Sendo assim, nesta seção serão realizada a equivalência pela descrição de cada uma das métricas.

As medidas de desempenho são baseadas na quantidade de exemplos de teste classificados correta e incorretamente, sendo correto quando a classe predita é igual à classe verdadeira e incorreto o contrário. Para tanto, dado um classificador e um exemplo a ser classificado em duas classes possíveis, as quatro possibilidades de saída do classificador são:

verdadeiro positivo: exemplo com rótulo ”positivo” classificado com classe ”positivo”

falso positivo: exemplo com rótulo ”negativo” classificado com classe ”positivo”

verdadeiro negativo: exemplo com rótulo ”negativo” classificado com classe ”negativo”

falso negativo: exemplo com rótulo ”positivo” classificado com classe ”negativo”

Ao submeter um conjunto de exemplos de teste ao classificador, é gerada uma saída para cada exemplo. A quantidade final de Verdadeiros Positivos

(VP), Falsos Positivos (FP), Verdadeiros Negativos (VN) e Falsos negativos (FN) pode ser organizada em uma matriz de confusão (ou matriz de contingência). A Tabela 2.3 mostra a disposição de VP, FP, VN e FN na matriz de confusão.

	preditos positivo	preditos negativo	
exemplos positivos	*VP*	*FN*	*Pos*
exemplos negativos	*FP*	*VN*	*Neg*
	PPos	*PNeg*	*Total*

Tabela 2.3: Medidas que compõem a matriz de confusão

Observe que a diagonal principal da matriz representa as predições corretas por classe, enquanto a diagonal secundária indica as classificações incorretas por classe. Duas medidas que podem ser extraídas da matriz, inicialmente, são as Taxas de Verdadeiros Positivos (TVP) e de Falsos Positivos (TFP). As fórmulas de cálculo de TVP e TFP podem ser vistas nas equações 2.5 e 2.6.

$$TVP = \frac{VP}{Pos} \tag{2.5}$$

$$TFP = \frac{FP}{Neg} \tag{2.6}$$

Além das taxas TVP e TFP, as medidas de precisão, acurácia e erro também podem ser calculadas com os valores da matriz de confusão. A descrição a seguir mostra a maneira de calcular cada uma dessas medidas.

acurácia (*precision*) : é a quantidade de exemplos verdadeiros classificados corretamente sobre o total de positivos. Por exemplo, é a quantidade de pacientes doentes diagnosticados doentes, sobre o total de pacientes doentes.

$$Acurácia = \frac{VP}{PPos} \tag{2.7}$$

precisão (*accuracy*) : é a quantidade de exemplos, positivos ou negativos, classificados corretamente, sobre o total de exemplos. Por exemplo, é a quantidade de pacientes doentes diagnosticados doentes, mais a quantidade de pacientes não doentes diagnosticados não doentes, sobre o total de pacientes.

$$Precisão = \frac{VP + VN}{Total} \tag{2.8}$$

erro : é o complemento da precisão

$$Erro = 1 - Precisão \tag{2.9}$$

revocação (*recall*) : é um sinônimo da taxa de verdadeiros positivos (TVP) descrita acima em 2.5.

$$Revogacão = TVP = \frac{VP}{Pos} \tag{2.10}$$

Um ponto relevante quanto às medidas de avaliação de classificadores descritas, é que algumas são diretamente relacionadas à distribuição de classes dos dados. A medida de acurácia é um exemplo disso, quando utilizada com classes desbalanceadas, pode "esconder" o erro de predição da classe minoritária. Isso implica em uma avaliação de desempenho tendenciosa. Para amenizar esse problema, pode-se comparar a precisão dos resultados com a precisão da classe majoritária, e considerar relevante somente os resultados maiores que o desempenho da classe majoritária.

Análise ROC

Outra maneira de avaliar os classificadores, e que também leva em consideração o desbalanceamento das classes, é a análise ROC, proposta por Fawcett (2006) para avaliação de desempenho e calibração de classificadores.

Trata-se de uma maneira de avaliar o comportamento do classificador utilizando um gráfico, denominado ROC, que permite a visualização de diferentes perspectivas da classificação.

A Figura 2.4 mostra um exemplo do gráfico ROC. Observe que os eixos horizontal e vertical variam entre 0 e 1 e são obtidos de TFP e TVP, respectivamente. Os pontos que formam o gráfico são tuplas (TFP,TVP), que representam o mesmo classificador sob diferentes perspectivas, uma vez que TFP e TVP dependem do resultado de classificação e, consequentemente, do limiar de decisão adotado pelo classificador.

As áreas de Ceu e Inferno ROC indicadas na Figura 2.4 representam a predominância de verdadeiros positivos e falsos positivos, respectivamente. O ponto (0,1), no canto superior esquerdo, e o ponto (1,0), no canto inferior direito, resultam da classificação correta todos os exemplos positivos e negativos, respectivamente. Seguindo o mesmo raciocínio, no Ceu ROC os verdadeiros positivos são quase sempre predominantes e, no Inferno ROC os verdadeiros negativos são quase sempre predominantes.

A linha na diagonal principal da Figura 2.4 representa o gráfico ROC de um classificador aleatório, que atribui classes valores de confiança indistintamente aos exemplos. O desenho do gráfico de um classificador é obtido pelo ranqueamento do valor de confiança predito e pela classe verdadeira de cada

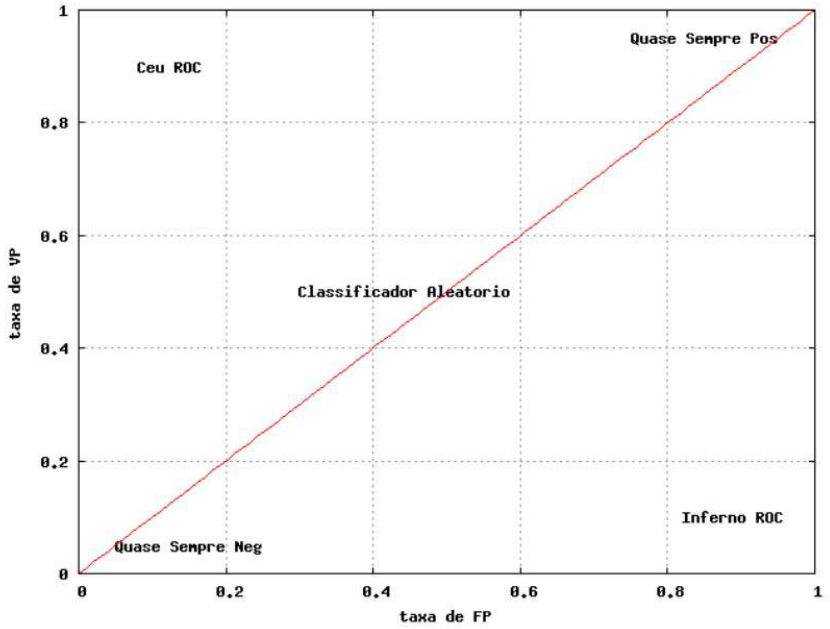

Figura 2.4: Gráfico ROC de um classificador aleatório

exemplo. O valor de confiança utilizado como *rank* é o predito pelo classificador para a classe positiva. A Tabela 2.4 mostra um exemplo de ordenação por *rank* com 20 exemplos. Na primeira coluna tem-se o número do exemplo, seguido pela confiança (*rank*), pela classe verdadeira do exemplo e, por fim, pela direção do desenho na curva ROC.

A lógica de interpretação da ordenação por *rank* e da classificação, o limiar de decisão que separa os exemplos preditos em positivos e negativos é representado por um valor de confiança na ordenação. Os exemplos com valor de confiança menor e maior que o limiar são preditos como negativos e positivos, respectivamente. Assim, quando um exemplo negativo recebe um valor de confiança maior que o valor de confiança de um exemplo positivo, ele se perde entre os positivos, ou seja, é "predito" como positivo e é contabilizado como um erro para o classificador. Na Tabela 2.4 por exemplo, considerando um limiar de 0.50, o exemplo da linha 12 é negativo e tem valor de confiança de 0.55, que é maior que o limiar de decisão e situa o exemplo 12 entre os positivos.

Seguindo a mesma lógica de interpretação, partindo do ponto (0,0) do gráfico e do primeiro exemplo da ordenação ilustrada na Tabela 2.4, os exemplos preditos como positivo (valor de confiança menor que o limiar de decisão) fazem o desenho do gráfico subir, por isso na última coluna tem-se a direção "cima". Por outo lado, os exemplos preditos como negativos (valor de confiança menor que o limiar de decisão) fazem o desenho do gráfico seguir para a "direita". A Figura 2.5 mostra a curva para a ordenação da Tabela 2.4. Observe que os deslocamentos para a "direita" ocorrem nos exemplos $7, 10, 12 - 13$ e, os deslocamentos para cima, nos exemplos de $1 - 6, 8, 9, 11$.

Além de verificar visualmente o quanto o gráfico se aproxima das regiões

Exemplo	Pontuação (valor de confiança)	Classe verdadeira	Direção na curva ROC
1	0.99	positiva	cima
2	0.91	positiva	cima
3	0.9	positiva	cima
4	0.88	positiva	cima
5	0.8	positiva	cima
6	0.77	positiva	cima
7	0.75	negativa	direita
8	0.67	positiva	cima
9	0.67	positiva	cima
10	0.62	negativa	direita
11	0.59	positiva	cima
12	0.51	negativa	direita
13	0.5	negativa	direita
14	0.45	negativa	direita
15	0.4	negativa	direita
16	0.38	negativa	direita
17	0.3	negativa	direita
18	0.2	negativa	direita
19	0.18	negativa	direita
20	0.11	negativa	direita

Tabela 2.4: Ordenação por *rank* dos valores de confiança atribuídos por um classificador a um conjunto de exemplos rotulados e as respectivas direções na curva ROC resultantes das combinações de ordenação e classe verdadeira

de Céu e Inferno ROC, é interessante analisar o tamanho da área abaixo da curva (AUC). Quanto mais próximo ao Céu ROC, maior a área abaixo da curva e, consequentemente, melhor o classificador na separação dos exemplos entre as classes. Nesse sentido, utiliza-se a medida de Área Abaixo da Curva (AUC), que mede justamente o tamanho da área abaixo da curva do gráfico ROC, que varia entre 0 e 1.

Sabe-se que o valor ideal de AUC é 1, quando o gráfico alcança o CÉU ROC e todos os exemplos positivos tem valor de confiança maior que os exemplos negativos. Entretanto, o ideal nem sempre é atingido e, eventualmente, alguns ou muitos exemplos são classificados incorretamente. Nesse caso, são definidas estratégias de balanceamento entre as classes, de acordo com a aplicação, e aplica-se calibração de limiares de decisão para aumentar o desempenho da classificação.

Calibração de limiares de decisão

Como descrito anteriormente nesta seção, o limiar de decisão é um valor de confiança que divide os exemplos apresentados ao classificador em positivos e negativos. Sabe-se que a medida de desempenho da classificação depende do propósito do problema para o qual o classificador é induzido. Dessa maneira, o valor ideal para o limiar depende da medida de desempenho que se

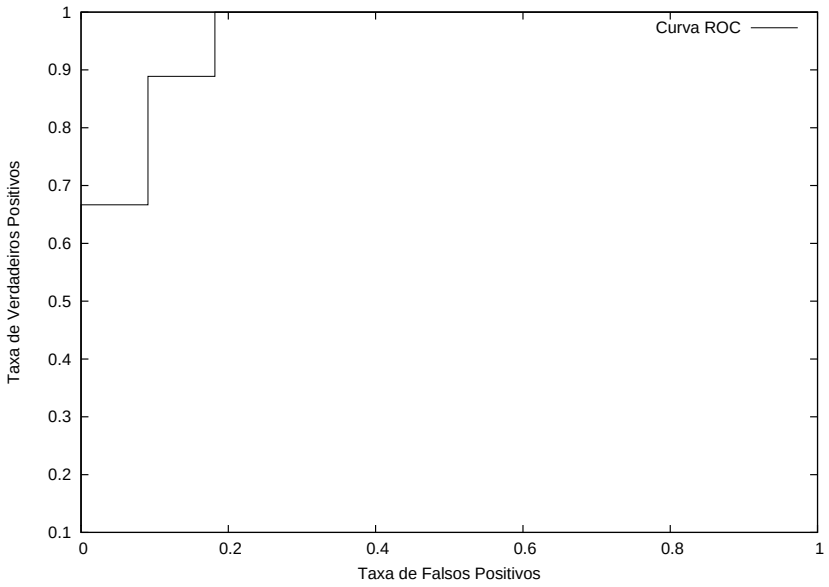

Figura 2.5: Curva ROC do classificador resultante da combinação dos valores de confiança ordenados por *rank* da Tabela 2.4

deseja otimizar. Nesse sentido utiliza-se a técnica de calibração de limiares de decisão, que consiste na busca por valores de confiança que otimizam determinada medida de desempenho.

Na ordenação da Tabela 2.4, por exemplo, a calibração do limiar de decisão para minimizar a taxa de erro poderia resultar em 0.67. Assim, antes da calibração, quando o limiar padrão de 0.50 é utilizado, 4 de 20 exemplos são classificados incorretamente, como pode ser visto na Tabela 2.5 nas linhas em negrito. Em 2.6 é ilustrada a matriz de confusão resultante da classificação dos exemplos da ordenação por *rank* da Tabela 2.5. Nas Equações 2.11 e 2.12 também são ilustradas as medidas de acurácia e precisão, com os valores 0.69 e 0.80, respectivamente.

$$Acurácia_{antes} = \frac{9}{13} = 0.69 \tag{2.11}$$

$$Precisão_{antes} = \frac{9+7}{20} = \frac{16}{20} = 0.80 \tag{2.12}$$

Alterando o limiar para 0.67, o número de exemplos classificados incorretamente diminui para 2, como ilustrado na Tabela 2.7 pelas linhas em negrito. A matriz de confusão também é recalculada e mostrada na Figura 2.8 e, logo abaixo, os valores de precisão e acurácia de 0.89 e 0.90, respectivamente. Observe que, com apenas a calibração do limiar de decisão, o mesmo classificador passou de uma precisão de 0.69 para 0.89 (Equação 2.14) e uma acurácia de 0.80 para 0.90 (2.13).

$$Acurácia_{depois} = \frac{8}{9} = 0.89 \tag{2.13}$$

Exemplo	Pontuação (valor de confiança)	Classe verdadeira	Classe predita
1	0.99	positiva	positiva
2	0.91	positiva	positiva
3	0.9	positiva	positiva
4	0.88	positiva	positiva
5	0.8	positiva	positiva
6	0.77	positiva	positiva
7	**0.75**	**negativa**	**positiva**
8	0.67	positiva	positiva
9	0.67	positiva	positiva
10	**0.62**	**negativa**	**positiva**
11	0.59	positiva	positiva
12	**0.51**	**negativa**	**positiva**
13	**0.5**	**negativa**	**positiva**
14	0.45	negativa	negativa
15	0.4	negativa	negativa
16	0.38	negativa	negativa
17	0.3	negativa	negativa
18	0.2	negativa	negativa
19	0.18	negativa	negativa
20	0.11	negativa	negativa

Tabela 2.5: Classificação dos valores de confiança da Tabela 2.4 em positivos e negativos usando um limiar de classificação padrão de 0.50

	preditos positivo	preditos negativo	
exemplos positivos	*9*	*0*	*9*
exemplos negativos	*4*	*7*	*11*
	13	*7*	*20*

Tabela 2.6: Matriz de confusão resultante da classificação dos exemplos da Tabela 2.5 (antes da calibração)

$$Precisão_{depois} = \frac{8+10}{20} = \frac{18}{20} = 0.90 \tag{2.14}$$

Outro exemplo de calibração pode buscar maximizar a acurácia de classificação. Para uma aplicação de filtro de spam, por exemplo, pode-se maximizar a acurácia da classe de e-mails importantes. É mais interessante que o usuário tenha alguns *spam* na caixa de entrada e não perca os e-mails importantes, do que se ver livre de *spam* sob a ameaça de perder e-mails importantes. Nesse sentido, considere que a Tabela 2.4 tenha como classe positiva os e-mails importantes e como classe negativa os *spam*.

A calibração para o problema de *spam* busca classificar corretamente todos os exemplos positivos (maior acurácia), independente de predizer alguns negativos incorretamente. Observe que não se trata de errar menos, ou de

Exemplo	Pontuação (valor de confiança)	Classe verdadeira	Classe predita
1	0.99	positiva	positiva
2	0.91	positiva	positiva
3	0.9	positiva	positiva
4	0.88	positiva	positiva
5	0.8	positiva	positiva
6	0.77	positiva	positiva
7	**0.75**	**negativa**	**positiva**
8	0.67	positiva	positiva
9	0.67	positiva	positiva
10	0.62	negativa	negativa
11	**0.59**	**positiva**	**negativa**
12	0.51	negativa	negativa
13	0.5	negativa	negativa
14	0.45	negativa	negativa
15	0.4	negativa	negativa
16	0.38	negativa	negativa
17	0.3	negativa	negativa
18	0.2	negativa	negativa
19	0.18	negativa	negativa
20	0.11	negativa	negativa

Tabela 2.7: Classificação dos valores de confiança da Tabela 2.4 usando um limiar de classificação calibrado em 0.67 para minimizar taxa de erro (depois da calibração)

	preditos positivo	preditos negativo	
exemplos positivos	*8*	*1*	*9*
exemplos negativos	*1*	*10*	*11*
	9	*11*	*20*

Tabela 2.8: Matriz de confusão resultante da classificação dos exemplos da Tabela 2.7 (depois da calibração)

levar em consideração o desempenho das duas classes, mais sim de acertar o máximo possível de uma determinada classe (positiva).

Nesse sentido, o valor de confiança 0.59 é o mais indicado para o limiar de decisão do problema de *spam*, porque com ele é possível predizer todos os exemplos positivos corretamente, alcançando uma acurácia de 1.0. Na Tabela 2.10 é ilustrada a separação para o problema de filtro de spam com o limiar calibrado. Na Tabela 2.11 é ilustrada a matriz de confusão e, em seguida, as medidas de acurácia (Equação 2.15) e precisão (Equação 2.16) recalculadas.

Na Tabela 2.9 são ilustradas as medidas de precisão e acurácia antes e depois das duas calibrações de limiar.

$$Acurácia_{spam} = \frac{9}{9} = 1.0 \tag{2.15}$$

Medida	Antes da calibração	1_a Calibração (taxa de erro)	2_a Calibração (*spam*)
Precisão	0.80	0.90	0.90
Acurácia	0.69	0.89	1.0

Tabela 2.9: Precisão e acurácia antes e depois das calibrações de limiar de decisão do classificador

Exemplo	Pontuação (valor de confiança)	Classe verdadeira	Classe predita
1	0.99	positiva	positiva
2	0.91	positiva	positiva
3	0.9	positiva	positiva
4	0.88	positiva	positiva
5	0.8	positiva	positiva
6	0.77	positiva	positiva
7	**0.75**	**negativa**	**positiva**
8	0.67	positiva	positiva
9	0.67	positiva	positiva
10	**0.62**	**negativa**	**positiva**
11	0.59	positiva	positiva
12	0.51	negativa	negativa
13	0.5	negativa	negativa
14	0.45	negativa	negativa
15	0.4	negativa	negativa
16	0.38	negativa	negativa
17	0.3	negativa	negativa
18	0.2	negativa	negativa
19	0.18	negativa	negativa
20	0.11	negativa	negativa

Tabela 2.10: Classificação dos valores de confiança da Tabela 2.4 usando um limiar de classificação calibrado em 0.59 para o problema de filtro de *spam*

	preditos positivo	preditos negativo	
exemplos positivos	*9*	*0*	*9*
exemplos negativos	*2*	*9*	*11*
	11	*9*	*20*

Tabela 2.11: Matriz de confusão resultante da classificação dos exemplos da Tabela 2.10 (depois da calibração para filtro de *spam*)

$$Precisão_{spam} = \frac{9+9}{20} = \frac{18}{20} = 0.90 \tag{2.16}$$

A calibração também pode ser obtida a partir da curva ROC. Cada ponto da curva representa uma Taxa de Verdadeiros Positivos (TVP), uma Taxa de Falsos Positivos (TFP), bem como todas as demais medidas de desempenho de classificação. Nesse sentido, os pontos podem ser considerados limiares de decisão que, ao serem escolhidos para a classificação, geram as medidas de desempenho a ele associadas (derivadas de VP e FP).

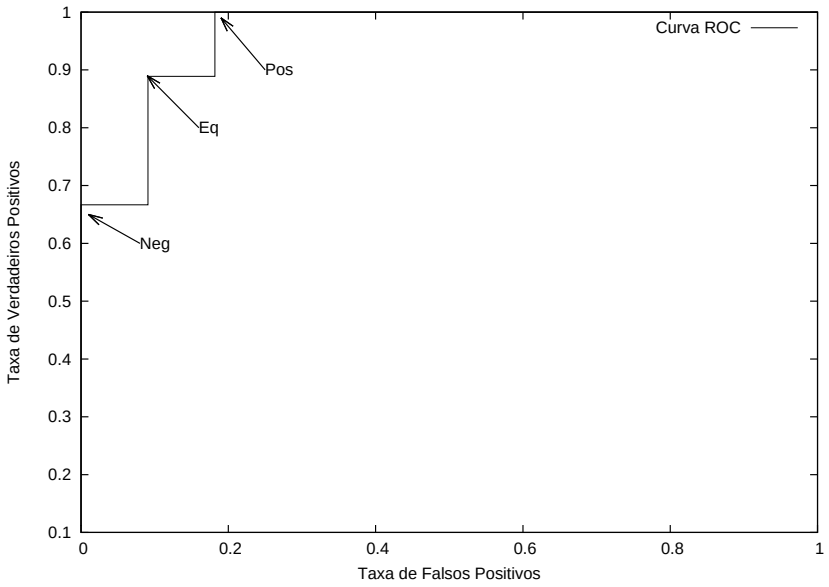

Figura 2.6: Curva ROC com opções de limiares de decisão para minimizar a taxa de erro de exemplos positivos, negativos e para manter um equilíbrio entre ambos

Visualmente, os pontos mais próximos ao eixo Y (mais afastados do Inferno ROC) geram classificadores mais conservadores da pureza de classificação da classe positiva, como ilustrado na Figura 2.6 pelo ponto marcado com 'Neg'. Observe que o ponto 'Neg' usado como limiar de decisão diminui a quantidade de exemplos negativos classificados incorretamente e buscam garantir que os exemplos preditos positivos sejam realmente positivos. Por outro lado, o ponto mais próximo à área superior do gráfico, ilustrado pela letra 'Pos', quando selecionado como limiar de decisão torna o classificador menos conservador para a classe positiva e busca garantir que todos os exemplos positivos sejam classificados corretamente. Por fim, o ponto mais próximo ao canto superior esquerdo (coordenadas 0 e 1), marcado com 'Eq' na figura, são limiares que otimizam a taxa de acerto geral do classificador, uma vez que representam um equilíbrio entre o Céu e o Inferno ROC.

Por fim, pode-se concluir que a partir da curva ROC é possível avaliar o desempenho de classificação, independente da distribuição de classes do problema. Também é possível identificar limiares de decisão que otimizam a medida de desempenho adequada para o problema de classificação.

CAPÍTULO

3

Mercado financeiro

Na área de predição de ações utilizando aprendizado de máquina, as representações de dados são constituídas por informações quantitativas históricas das ações. As Seções 3.0.5, 3.0.6, 3.0.7 e 3.0.8 explicam os conceitos de cotações de ações, índices de mercado e indicadores econômicos.

3.0.5 Cotações

Neste trabalho, as *cotações* são consideradas preços de mercado atribuídos a uma ação. Por meio do histórico de cotações, podem ser derivadas diversas outras métricas, medidas e informações para representar o comportamento da ação no mercado financeiro. O gráfico da Figura 3.1 é composto pelas cotações de fechamento diárias da ação PETR4 da PETROBRÁS, no período compreendido entre maio e julho de 2013. O dia 20 de julho está marcado com a letra D (*Decisão*), ressaltando-o como o dia que se deseja prever o comportamento. O período anterior ao dia D, denominado H, é representado por uma linha tracejada que mostra o comportamento passado da ação PETR4.

Com base no histórico da ação, as cotações imediatamente anteriores ao dia D são usadas para predizer o comportamento do preço da ação no dia D. No Gráfico da Figura 3.1, o período H está entre 2 de maio e 19 de julho e é o período utilizado para predizer o dia 20 de julho. Observe que, embora existam oscilações bruscas de preço no período H, a tendência mais recente ao dia D é de subida. Essa tendência histórica, em conjunto com outras características da ação, podem indicar a continuidade da tendência no futuro.

Outra característica que pode ser extraída do período H é a existência de padrões de comportamento. A Figura 3.2 mostra o mesmo intervalo de preços

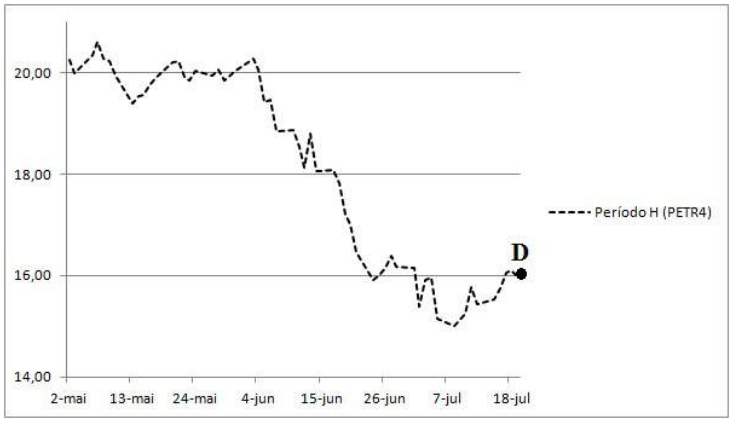

Figura 3.1: Preços de fechamento diários da PETR4 no período de maio a julho de 2013, com o dia D e período H identificados.

da PETR4, com duas sequências de cotações circuladas. Entre 7 e 13 de junho e entre 28 de junho e 4 de julho, a ação se comporta da mesma forma: estável, cai e sobe. Além da sequência ser a mesma, as proporções de variação são similares, ocasionando a repetição do formato do gráfico nas duas situações. Essa repetição de uma mesma sequência de cotações pode ser um padrão de comportamento de preço útil na predição da ação.

Dessa forma, a análise das cotações históricas pode auxiliar na predição de tendência futura da ação, possibilitando a extração de padrões de comportamento de preço. Tais informações podem ser utilizadas como atributos nas representações de dados.

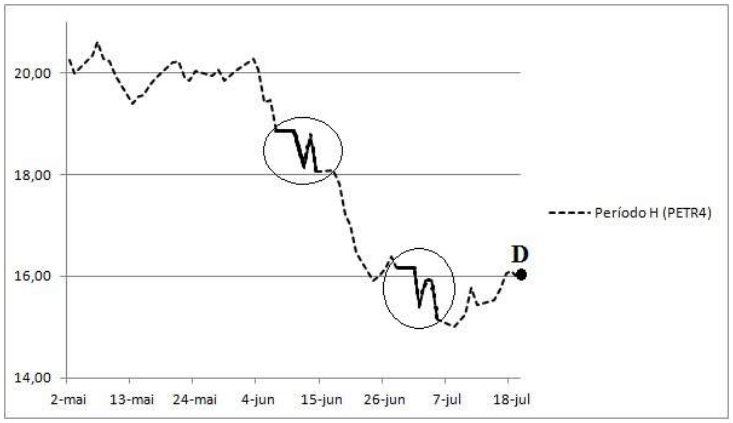

Figura 3.2: Preços de fechamento diários da PETR4 no período de maio a julho de 2013, com um padrão de cotações tracejado

3.0.6 Índices

Os índices agregam informações das ações mais negociadas no mercado. Os mais comuns são os índices das bolsas de valores, que indicam o movimento dos preços do mercado como um todo. O índice IBOVESPA, por exemplo, é composto pelas cotações das ações mais negociadas da Bolsa de Valores

de São Paulo. A Tabela 3.1 mostra as ações que compõe o IBOVESPA no mês maio de 2014. A primeira coluna é o código, seguido do nome da ação e, por último, o percentual de participação que a ação tem no cálculo do índice. Note que ações de empresas como Petrobrás e Vale contribuem notoriamente para o índice, com uma participação entre 7 e 8%. Enquanto outras como Usiminas e V-agro correspondem a aproximadamente 0,2% do índice.

Código	Ação	Participação(%)	Código	Ação	Participação(%)
ABEV3	AMBEV S/A	5.907	GGBR4	GERDAU	1.321
AEDU3	ANHANGUERA	0.773	GOAU4	GERDAU MET	0.501
ALLL3	ALL AMER LAT	0.4	GOLL4	GOL	0.156
BBAS3	BRASIL	2.236	HGTX3	CIA HERING	0.318
BBDC3	BRADESCO	1.651	HYPE3	HYPERMARCAS	0.676
BBDC4	BRADESCO	7.238	ITSA4	ITAUSA	2.978
BBSE3	BBSEGURIDADE	2.097	ITUB4	ITAUUNIBANCO	9.625
BISA3	BROOKFIELD	0.044	JBSS3	JBS	1.338
BRAP4	BRADESPAR	0.478	KLBN11	KLABIN S/A	0.581
BRFS3	BRF SA	3.385	KROT3	KROTON	1.273
BRKM5	BRASKEM	0.43	LAME4	LOJAS AMERIC	0.635
BRML3	BR MALLS PAR	0.886	LIGT3	LIGHT S/A	0.217
BRPR3	BR PROPERT	0.627	LREN3	LOJAS RENNER	0.94
BVMF3	BMFBOVESPA	2.336	MMXM3	MMX MINER	0.019
CCRO3	CCR SA	1.689	MRFG3	MARFRIG	0.196
CESP6	CESP	0.544	MRVE3	MRV	0.238
CIEL3	CIELO	2.985	NATU3	NATURA	0.739
CMIG4	CEMIG	1.35	OIBR4	OI	0.63
CPFE3	CPFL ENERGIA	0.595	PCAR4	P.ACUCAR-CBD	1.814
CPLE6	COPEL	0.37	PDGR3	PDG REALT	0.228
CRUZ3	SOUZA CRUZ	0.99	PETR3	PETROBRAS	4.871
CSAN3	COSAN	0.437	PETR4	PETROBRAS	7.736
CSNA3	SID NACIONAL	0.685	QUAL3	QUALICORP	0.514
CTIP3	CETIP	0.814	RENT3	LOCALIZA	0.593
CYRE3	CYRELA REALT	0.271	RSID3	ROSSI RESID	0.067
DTEX3	DURATEX	0.243	SANB11	SANTANDER BR	1.548
ECOR3	ECORODOVIAS	0.319	SBSP3	SABESP	0.831
ELET3	ELETROBRAS	0.165	SUZB5	SUZANO PAPEL	0.407
ELET6	ELETROBRAS	0.237	TBLE3	TRACTEBEL	0.775
ELPL4	ELETROPAULO	0.093	TIMP3	TIM PART S/A	1.072
EMBR3	EMBRAER	1.619	UGPA3	ULTRAPAR	1.97
ENBR3	ENERGIAS BR	0.243	USIM5	USIMINAS	0.431
ESTC3	ESTACIO PART	0.861	VALE3	VALE	4.106
EVEN3	EVEN	0.161	VALE5	VALE	5.547
FIBR3	FIBRIA	0.515	VIVT4	TELEF BRASIL	1.318
GFSA3	GAFISA	0.117	-	-	-

Tabela 3.1: Ações que compõem o índice IBOVESPA e suas respectivas porcentagens de participação

O IBOVESPA é calculado com base no preço das ações que o compõe e no percentual de participação de cada ação. A Figura 3.3 mostra os valores do índice entre os meses de maio e julho de 2013. Observe que a tendência mais geral é de queda, indicando que as cotações das ações da Tabela 3.1 estiveram em tendência de descida no período entre final de maio e começo de julho de 2013.

A utilização de índices em representações de dados pode partir do mesmo pressuposto das cotações: identificação de tendência e padrões de comportamento de preço. Entretanto, os índices tem a característica de oferecer um panorama médio sobre as cotações de várias ações. Sendo assim, a influência do índice não é específica para uma ação, mas sim para o ambiente que se deseja prever como um todo. Nesse sentido, é preciso identificar como cada ação reage ao valor no índice.

Considere que se deseja verificar a relação entre a ação PETR4 e o índice IBOVESPA da Figura 3.3. Note a similaridade entre os gráficos, e perceba como o IBOVESPA poderia ser utilizado para predizer a PETR4. A Figura 3.3

sinaliza uma queda geral dos preços no período H, pois o IBOVESPA está em tendência de descida. Tal observação, em conjunto com outras específicas da ação PETR4, podem indicar uma tendência de descida também para o preço da PETR4.

Embora a relação da PETR4 com o IBOVESPA seja privilegiada, por conta da participação de 7% no valor do índice, a predição das demais ações também pode usufruir do humor do mercado que o IBOVESPA pode fornecer. Cada qual no seu modo, as ações podem aliar a tendência mais forte que o índice oferece com informações específicas da ação que se deseja prever, como cotações e indicadores econômicos.

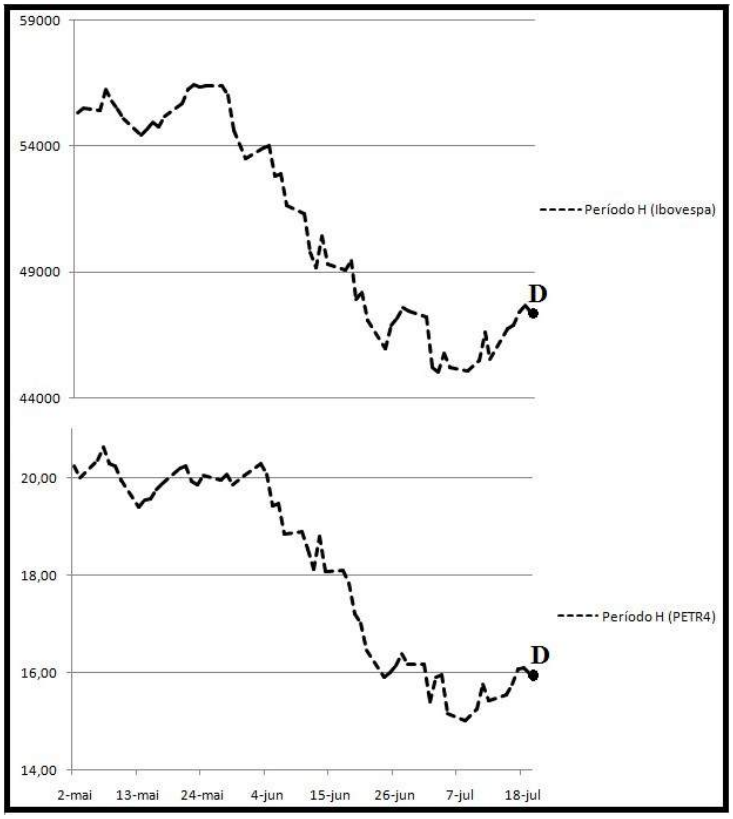

Figura 3.3: Valores do índice IBOVESPA entre os meses de maio e julho de 2013, com o dia D e período H identificados

3.0.7 Sobrecompra e sobrevenda

A maioria dos indicadores econômicos descritos na Seção 3.0.8, utiliza os conceitos de sobrecompra e sobrevenda na interpretação de seus valores. Ambos os termos referem-se a limites que, quando ultrapassos, indicam operações de compra e venda. A seguir são descritos os detalhes desses limites e, na Figura 3.4 é mostrado um exemplo da aplicação dos conceitos de sobrecompra e sobrevenda na interpretação de um indicador qualquer.

Sobrecompra: o termo sobrecompra refere-se à situação na qual os preços da

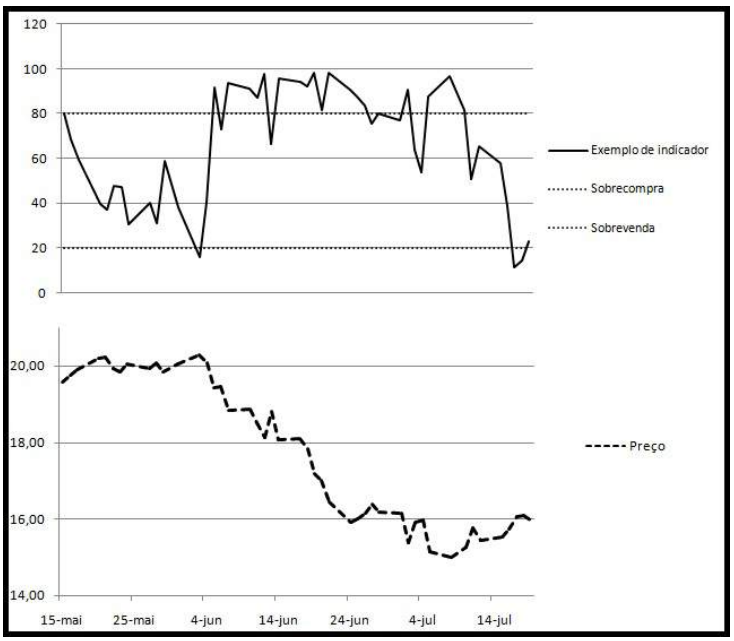

Figura 3.4: Exemplo de indicador econômico e série temporal de preço com os índices de sobrecompra e sobrevenda sinalizados

ação passam por um processo de alta prolongada (ou intensa) e tendem a se desvalorizar. No gráfico dos indicadores esse comportamento pode ser visualizado quando o valor do indicador ultrapassa o limites superiores de sobrecompra. Em geral, os limites são estabelecidos entre 60 e 90 dependendo da tendência da ação, sendo 60 para tendência de baixa e 90 para de alta. Na Figura 3.4, a área superior do gráfico do indicador, acima de 80, é denominada área de sobrevenda. Quando o valor do indicador entra na área de sobrevenda, acredita-se na possibilidade do movimento do preço reverter para descida. Observe que, de fato, no período entre 04/junho e 04/julho, o indicador ultrapassa o limite superior, sinalizando a sobrevenda, e a série temporal de preço (ilustrada na parte inferior da figura) entra em em movimento de descida. Assim, o limite de sobrecompra é usado para sinalizar a venda das ações, buscando evitar que o investidor tenha prejuízos com a ação cujo preço tende a desvalorizar.

Sobrevenda: a sobrevenda é o inverso da sobrecompra e representa a situação na qual a ação se desvaloriza por muito tempo (ou em grande quantidade) e depois tende a subir. Assim, os limites utilizados para limitar a região de sobrevenda estão na parte inferior do gráfico, entre 10 e 40 dependendo da tendência da ação. Na Figura 3.4 o limite de sobrevenda está em 30 e, em dois momentos, o indicador ultrapassa esse limite, o que sinaliza possíveis reversões do movimento do preço para subida. O trecho circulado *Sobe* mostra um exemplo no qual o indicador sinaliza a sobrevenda e a série temporal do preço da ação (ilustrada na parte in-

ferior da figura) de fato mostra uma reversão para tendência de subida. Portanto, os limites de sobrevenda podem ser utilizados para sinalizar operações de compra de ações, pois os preços tendem a subir quando a ação está nessa situação, gerando possível lucro para o investidor.

3.0.8 Indicadores

Outro recurso das representações de dados para predição de ações são os indicadores econômicos. Provenientes da análise técnica, são cálculos elaborados baseados no histórico das cotações das ações. Como foram pesquisados durante anos por especialistas da área econômica, os indicadores econômicos agregam informações específicas do comportamento das ações, úteis para analisar situações como tendências, sinais de reversão de preços e momentos ideais para operações

Neste projeto foram considerados alguns indicadores descritos em (Huang and Tsai, 2009), (Saffi, 2003), (jae Kim, 2003) e (Klassen, 2005), bem como os mais utilizados por especialistas técnicos econômicos. Ainda nesta seção os indicadores selecionados são detalhados e alguns conceitos da análise técnica são brevemente descritos, para auxiliar a interpretação dos indicadores.

Descrição e exemplos

Nesta seção são descritos, exemplificados e interpretados os seguintes indicadores econômicos: Média Móvel Simples (MMS), Média Móvel Exponencial (MME), Convergência/Divergência da Média Móvel (MACD), Índice de Força Relativa (IFR), Estocástico (K e D), Larry Williams (Larry), Momento (MO) e Taxa de Mudança (ROC), Linha Psicológica (PSY) e Disparidade (Disp).

A explicação de cada indicador se divide em duas etapas: descrição da utilidade do indicador e exemplos de aplicação da fórmula; e interpretação de resultados. Ambas as etapas utilizam exemplos de cálculo do indicador para facilitar a compreensão. Os dados utilizados nos exemplos são da ação PETR4 da BOVESPA, no período de 2 de maio e 19 de julho de 2013.

A Tabela 3.2 e o gráfico da Figura 3.1 mostram os dados utilizados no cálculo dos indicadores. Na tabela, as colunas representam: o preço de fechamento diário, o preço máximo diário, o preço mínimo diário, o dia em que o indicador começou a ser calculado e a quantidade de dias utilizados no cálculo do indicador. Os dois últimos valores, início do indicador e período n variam para cada indicador. Por exemplo, para as taxas momento e ROC com $n = 3$, os três primeiros dias do período são usados para calcular o primeiro valor desses indicadores, no quarto dia do período (7-mai). As médias móveis com $n = 10$ usam 10 dias anteriores, incluindo o dia para o qual está sendo

calculada a média. Assim, o primeiro valor da MMS e da MME é calculado no 10_o dia do período (15-mai).

O período n utilizado em cada indicador foi selecionado por sugestões da literatura e de especialistas da área de análise técnica.

Por fim, o dia atual C_t, utilizado nos exemplos de aplicação da fórmula dos indicadores, referem-se ao dia 19 de julho de 2013.

Dia	Preço	Máximo	Mínimo	Início do indicador	Período n
2-mai	20,25	20,63	19,85		
3-mai	20,00	20,66	19,89		
6-mai	20,35	20,37	19,82		
7-mai	20,62	20,74	20,28	Momento e ROC	3
8-mai	20,29	20,82	20,22	Disparity	5
9-mai	20,24	20,52	20,09		
10-mai	19,95	20,39	19,71		
13-mai	19,39	19,85	19,38		
14-mai	19,55	19,68	19,25		
15-mai	19,57	19,64	19,33	MMS, MME e Larry	10
16-mai	19,75	20,08	19,46	PSY	10
17-mai	19,89	20,05	19,79		
20-mai	20,20	20,25	19,56		
21-mai	20,24	20,38	19,98	K	14
22-mai	19,92	20,40	19,68	IFR	14
23-mai	19,86	19,86	19,48	D	3
24-mai	20,05	20,16	19,79		
27-mai	19,94	20,15	19,92		
28-mai	20,07	20,34	20,01		
29-mai	19,85	20,07	19,80		
31-mai	20,05	20,36	19,80		
3-jun	20,29	20,44	19,94		
4-jun	20,07	20,47	19,96		
5-jun	19,42	20,13	19,32		
6-jun	19,46	19,55	19,08		
7-jun	18,84	19,27	18,72		
10-jun	18,88	19,12	18,83		
11-jun	18,53	18,79	18,23		
12-jun	18,12	18,80	18,06		
13-jun	18,80	19,02	17,94		
14-jun	18,06	18,86	18,00		
17-jun	18,09	18,53	17,96		
18-jun	17,83	18,08	17,62		
19-jun	17,20	17,95	17,15	MACD	12, 26 e 9
20-jun	17,02	17,27	16,50		
21-jun	16,46	16,93	16,40		
24-jun	15,91	16,25	15,57		
25-jun	16,00	16,39	15,87		
26-jun	16,15	16,44	16,08		
27-jun	16,38	16,60	16,17		
28-jun	16,17	16,35	15,94		
1-jul	16,16	16,35	15,94		
2-jul	15,39	16,25	15,11		
3-jul	15,90	15,98	15,15		
4-jul	15,96	16,17	15,88		
5-jul	15,15	15,82	14,94		
8-jul	15,00	15,35	14,97		
10-jul	15,25	15,44	15,06		
11-jul	15,76	15,83	15,25		
12-jul	15,43	15,92	15,27		
15-jul	15,54	15,69	15,35		
16-jul	15,75	15,80	15,47		
17-jul	16,06	16,20	15,87		
18-jul	16,10	16,29	15,91		
19-jul	15,98	16,10	15,85		

Tabela 3.2: Informações financeiras históricas da PETR4 utilizados no cálculo dos indicadores

Veja a seguir a descrição detalhada dos indicadores econômicos considerados:

Média Móvel Simples MMS: média do preço de n dias anteriores. Possibilita o acompanhamento das oscilações de preço, reduzindo o efeito de ruídos. O gráfico da Figura 3.5 mostra a comparação do comportamento do preço e da MMS. Como pode ser visualizado, a MMS suaviza os movimentos do preço, retirando as variações mais agudas. Ela é dita móvel porque, à medida que um novo dia é acrescentado ao cálculo, o último é retirado, como uma janela que desliza pelos preços.

Fórmula:

$$MMS(n)_t = \frac{1}{n} \times \sum_{i=t-n+1}^{t} C_i \tag{3.1}$$

onde n é o tamanho do período, C_i é o preço no dia i e t é o dia atual.

Exemplo para n = 10:

$$\begin{aligned} MMS(10)_{19} &= \frac{1}{10} \times \sum_{i=19-10+1}^{19} C_i \\ MMS(10)_{19} &= \frac{1}{10} \times (15,15 + 15,00 + 15,25 + 15,76 + \\ &+ 15,43 + 15,54 + 15,75 + 16,60 + 16,10 + 15,98) \\ MMS(10)_{19} &= 15,60 \end{aligned}$$

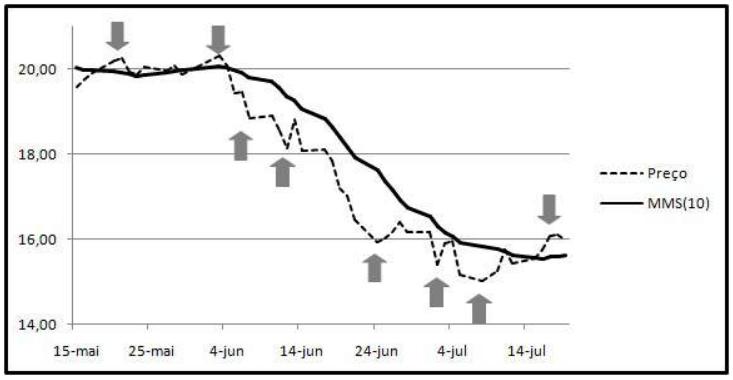

Figura 3.5: Média Móvel Simples e Preço da PETR4 no período de 15 de maio a 19 de julho de 2013

Média Móvel Exponencial: é uma média móvel de n dias anteriores assim como a MMS. Entretanto, atribui pesos maiores aos preços mais recentes, assim as oscilações de preço passam a ser visualizadas mais rapidamente. Consequentemente, o desenho do gráfico é menos suave.

Fórmula:

$$MME(n)_t = \frac{2}{n+1} C_t + \left(1 - \frac{2}{n+1}\right) \times MMS(n)_t \tag{3.2}$$

Exemplo para n = 10:

$$\begin{aligned} MME(10)_{19} &= \frac{2}{10+1} C_{19} + \left(1 - \frac{2}{10+1}\right) \times MMS(10)_{19} \\ MME(10)_{19} &= \left(\frac{2}{11} \times 15,98\right) + \left(1 - \frac{2}{11}\right) \times 15,6 \\ MME(10)_{19} &= 15,67 \end{aligned} \tag{3.3}$$

Interpretação da MMS e MME: a plotagem dos gráficos de preço e média móvel juntos permite a análise do comportamento do preço em relação à

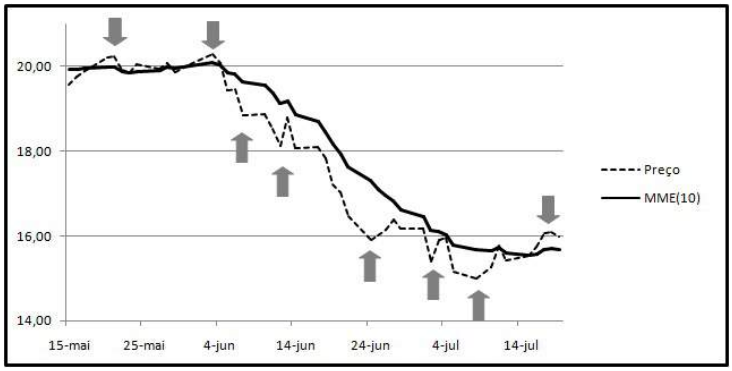

Figura 3.6: Média Móvel Exponencial e Preço da PETR4 no período de 15 de maio a 19 de julho de 2013

média do período no qual ele se encontra. Independente de ser MMS ou MME, à medida que o preço se afasta da média a ação pode entrar em situação de sobrecompra ou sobrevenda. Nas Figuras 3.5 e 3.6, as setas para baixo e para cima exemplificam os momentos nos quais os preços ultrapassam os limites de sobrecompra e sobrevenda, respectivamente. A diferença entre as linhas dos dois gráficos é quase imperceptível. Isso pode ter ocorrido porque a oscilação dos preços mais recentes foi pequena, e mesmo com os pesos a MME sofreu pouco influência.

Convergência/Divergência da Média Móvel de n dias: indicador de momento de compra e venda. Compara duas médias móveis para indicar a tendência e o momento ideal para operações de compra ou venda.

Fórmula:

$$MACD_t = MME(12)_t - MME(26)_t \tag{3.4}$$

A detecção do momento ideal é feita com a *linha de sinal*, que é o cálculo da MME(9) da MACD. A compra é indicada quando a MACD é maior que a linha de sinal, enquanto a venda quando a MACD é menor que a linha de sinal.

$$Signal \;=\; MME(9)_t(MACD) \tag{3.5}$$

Exemplo da MACD:

$$\begin{aligned} MACD_{19} &= MME(12)_{19} - MME(26)_{19} \\ MME(12)_{19} &= 15,70 \\ MME(26)_{19} &= 16,24 \\ MACD_{19} &= 15,70 - 16,24 \\ MACD_{19} &= -0,53 \end{aligned} \tag{3.6}$$

Exemplo da Linha de Sinal (Signal):

$$\begin{aligned} Signal &= MME(9)_{19} \\ Signal &= -0,88 \end{aligned} \tag{3.7}$$

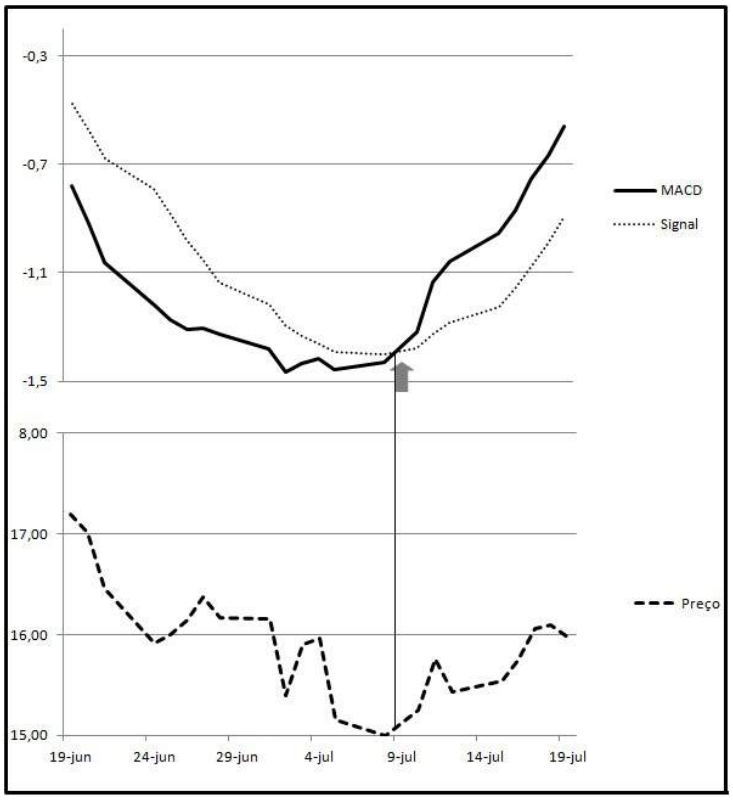

Figura 3.7: Convergência/Divergência da Média Móvel e Preço da PETR4 no período de 19 de junho a 19 de julho de 2013

Interpretação da MACD: a Figura 3.7 ilustra os gráficos de MACD e linha de sinal. O cruzamento das linhas, sinalizado com uma seta para cima, indica situação de sobrevenda porque a linha de MACD está acima da linha de sinal. Se, após o cruzamento, a linha de MACD ficasse abaixo da linha de sinal, a situação seria de sobrecompra.

Índice de força relativa: oscilador que acompanha o preço buscando identificar um potencial incremento da ação. O cálculo utiliza as médias de ganho e de perda, que são as médias de preço dos dias com tendência de subida e descida respectivamente. Os valores do indicador variam entre 0 e 100, onde valores próximos a 100 indicam mais ganho do que valores próximos a 0. Além disso, áreas de sobrecompra e sobrevenda são utilizadas na interpretação do indicador.

Fórmula:

$$IFR(n)_t = \left(100 - \left(\frac{100}{\left(1 + \frac{UPC_n/UD_n}{DPC_n/DD_n}\right)}\right)\right) \tag{3.8}$$

onde UD_n é a quantidade de dias com tendência de subida - ou *upward* - (durante os n dias anteriores), DD_n é a quantidade de dias com tendência de descida - *downward*, UPC_n é a soma dos preços nos dias de *upward* e DPC_n é a soma dos preços nos dias de *downward*.

Exemplo para n = 14:

$$
\begin{aligned}
IFR(14)_{19} &= \left(100 - \left(\frac{100}{\left(1 + \frac{UPC_{19}/UD_{19}}{DPC_{19}/DD_{19}}\right)}\right)\right) \\
IFR(14)_{19} &= \left(100 - \left(\frac{100}{\left(1 + \frac{\frac{15,90+15,96+15,25+15,76+15,54+15,75+16,06+16,10}{8}}{\frac{16,16+15,39+15,15+15+15,43+15,98}{6}}\right)}\right)\right) \\
IFR(14)_{19} &= \left(100 - \left(\frac{100}{\left(1 + \frac{15,79}{15,51}\right)}\right)\right) \\
IFR(14)_{19} &= 50,43
\end{aligned}
\tag{3.9}
$$

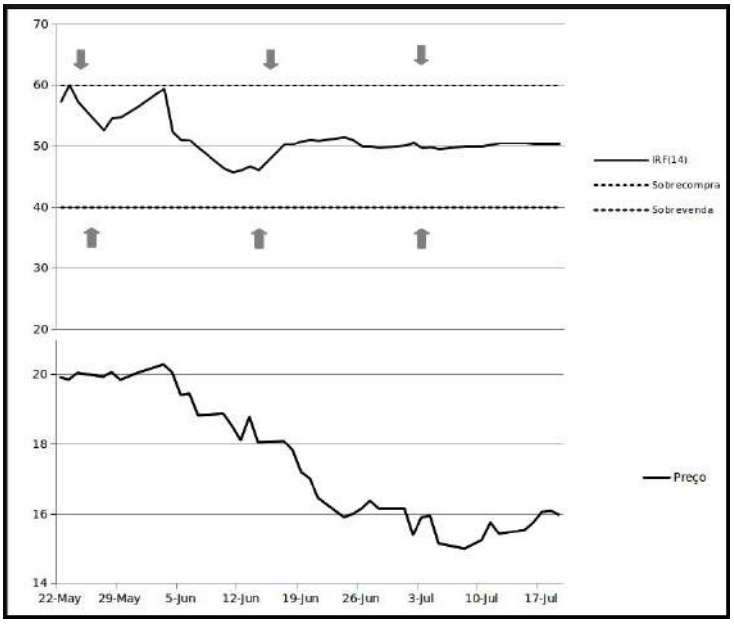

Figura 3.8: Índice de Força Relativa e Preço da PETR4 no período de 22 de maio a 19 de julho de 2013

Interpretação do IRF: o gráfico da Figura 3.8 mostra o indicador IFR e as áreas de sobrecompra e sobrevenda, sinalizadas com setas para baixo e para cima respectivamente. Mesmo com limites apertados - 40 e 60 - o indicador não ultrapassou os limites de sobrecompra e sobrevenda. Consequentemente pontos de reversão de preço não são detectados graficamente.

Estocástico: é composto pelos estocásticos K e D. Tenta determinar a relação entre o preço atual e os valores máximo e mínimo do período. O Esto-

cástico K relaciona o preço atual e a amplitude dos preços dos n dias anteriores. Sendo assim, é mais sensível às oscilações. Já o Estocástico D, suaviza o comportamento do K, também é conhecido por estocástico lento e se trata da Média Móvel Simples de K. A comparação dos valores de K e D mostra a tendência dos preços. No momento em que K se torna maior que D, tem-se uma tendência de subida. Quando, ao contrário, K se torna menor que D, a tendência é de descida. Além disso, áreas de sobrecompra e sobrevenda podem ser identificadas graficamente.

Fórmula do estocástico K:

$$K(n)_t = \frac{C_t - LP_n}{HP_n - LP_n} \times 100 \tag{3.10}$$

onde C_t é o preço no dia atual, LP_n e HP_n são a menor cotação mínima e a maior cotação máxima nos n dias anteriores, respectivamente.

Fórmula do estocástico D:

$$D(n)_t = \frac{\sum_{i=t-n+1}^{t} K(n)_i}{n} \tag{3.11}$$

Exemplo para n = 14:

$$\begin{aligned} K(14)_{19} &= \frac{C_{19} - LP_{14}}{HP_{14} - LP_{19-14}} \times 100 \\ K(14)_{19} &= \frac{15,98 - 14,94}{16,35 - 14,94} \times 100 \\ K(14)_{19} &= 73,75 \end{aligned} \tag{3.12}$$

$$\begin{aligned} D(14)_{19} &= \frac{\sum_{i=17}^{19} K(14)_i}{14} \\ D(14)_{19} &= \frac{K(14)_{17} + K(14)_{18} + K(14)_{19}}{3} \\ D(14)_{19} &= \frac{67,46 + 82,26 + 73,75}{3} \\ D(14)_{19} &= 74,49 \end{aligned} \tag{3.13}$$

Interpretação do K e D: A Figura 3.9 mostra o gráfico dos estocásticos K e D e as áreas de sobrecompra e sobrevenda, sinalizadas com setas para baixo e para cima. Com os limites de sobrecompra e sobrevenda ajustados em 70 e 10, a duas circunferências superiores sinalizam possíveis

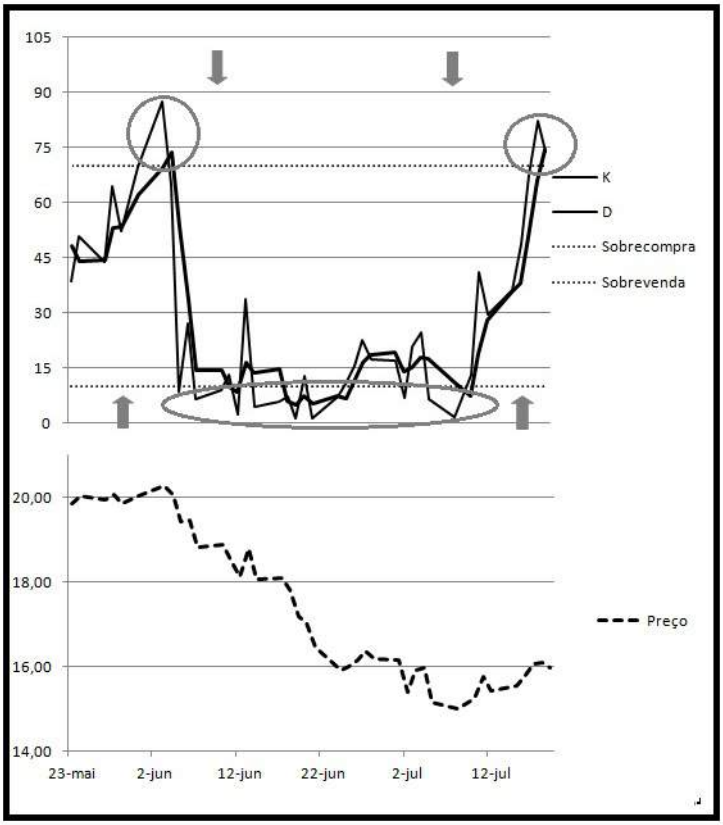

Figura 3.9: Indicadores estocásticos K e D e Preço da PETR4 no período de 23 de maio a 19 de julho de 2013

tendências de descida. Enquanto a circunferência inferior mostra uma série de pequenas tendências de subida. Além disso, a comparação das linhas K e D e do preço mostra que, em geral, quando a linha K está acima da D, o preço está subindo, e quando a D está acima da K, o preço está descendo.

Oscilador Larry Williams: em termos de cálculo é similar ao indicador estocástico, com a diferença de utilizar o máximo do período ao invés do mínimo. Também utiliza limites de sobrecompra e sobrevenda.

Fórmula:

$$Larry(n)_t = \frac{HP_n - C_t}{HP_n - LP_n} \times 100 \tag{3.14}$$

Exemplo para n = 10:

$$\begin{aligned} Larry(10)_{19} &= \frac{HP_{10} - C_{19}}{HP_{10} - LP_{10}} \times 100 \\ Larry(10)_{19} &= \frac{16,3 - 15,98}{16,3 - 14,94} \\ Larry(10)_{19} &= 22,96 \end{aligned} \tag{3.15}$$

Interpretação do Larry: O gráfico da Figura 3.10 mostra os valores do

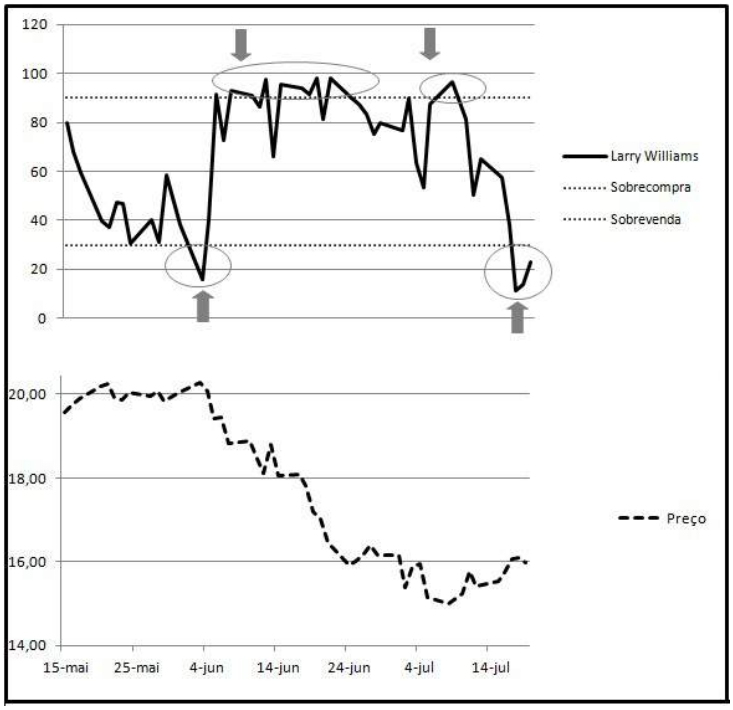

Figura 3.10: Oscilador Larry Williams e Preço da PETR4 no período de 15 de maio a 19 de julho de 2013

indicador Larry e as áreas de sobrecompra e sobrevenda, sinalizadas com as setas para baixo e para cima, respectivamente. A interpretação desse indicador é similar ao Estocástico. As circunferências superiores indicam situações de sobrecompra, enquanto as inferiores de sobrevenda.

Momento e Taxa de Mudança: medem o quanto o preço mudou ao longo de um período de tempo n. Valores altos ou baixos em relação aos demais do período podem indicar saturação da tendência e, consequentemente, níveis de sobrecompra e sobrevenda. A diferença é que os limites não são valores pré-definidos, como 30 e 70, porque esse indicador assume valores dentro de um intervalo desconhecido previamente. Então os limites de sobrecompra e sobrevenda são relativos ao intervalo de preços do período observado.

Fórmula do indicador Momento:

$$MO(n)_t = C_t - C_{t-n} \tag{3.16}$$

Fórmula do indicador Taxa de Mudança:

$$ROC(n)_t = \frac{C_t}{C_{t-n}} \times 100 \tag{3.17}$$

Exemplo para n = 3:

$$\begin{aligned} MO(3)_{19} &= C_{19} - C_{19-3} \\ MO(3)_{19} &= 15,98 - 15,75 \\ MO(3)_{19} &= 0,23 \end{aligned} \tag{3.18}$$

$$\begin{aligned} ROC(3)_{19} &= \frac{C_{19}}{C_{19-3}} \times 100 \\ ROC(3)_{19} &= \frac{15,98}{15,75} \\ ROC(3)_{19} &= 101,46 \end{aligned} \tag{3.19}$$

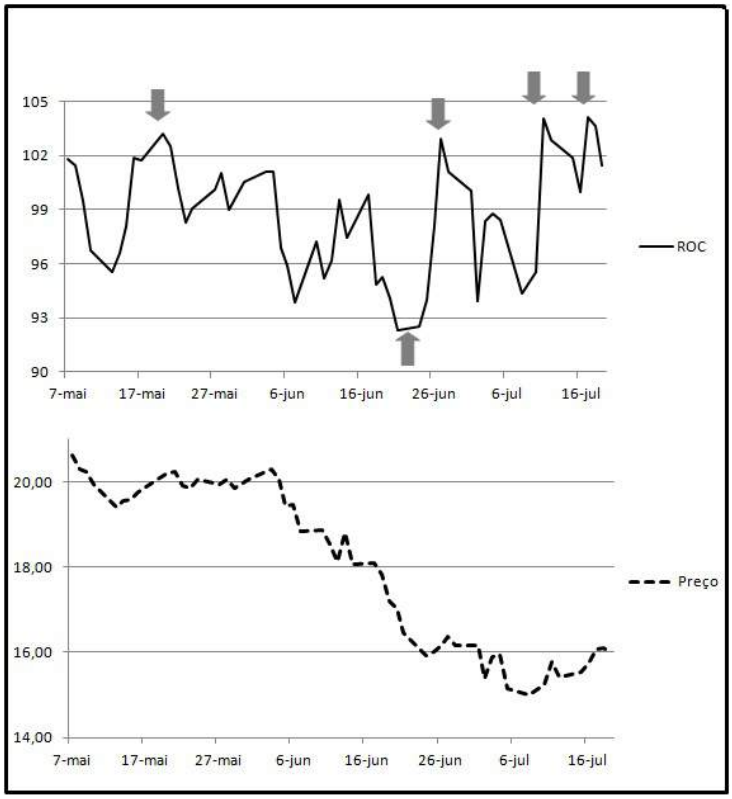

Figura 3.11: Indicador de Taxa de mudança e Preço da PETR4 no período de 7 de maio a 19 de julho de 2013

Interpretação do Momento e ROC: Os gráficos das Figuras 3.11 e 3.12 mostram o comportamento dos indicadores Taxa de Mudança e Momento. Como pode-se visualizar, o desenho de ambos os gráficos é bastante similar, diferenciando o intervalo de valores de cada um. Nos dois gráficos,

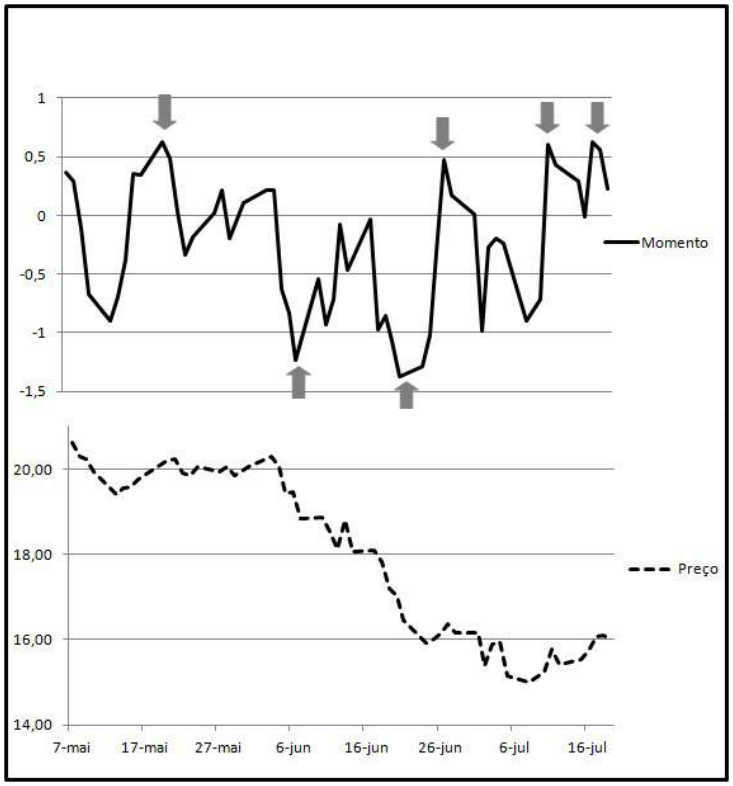

Figura 3.12: Indicador de Momento e Preço da PETR4 no período de 7 de maio a 19 de julho de 2013

os valores que mais destoam dos demais são sinalizados com setas: para baixo indicando sobrecompra e para cima sobrevenda. Como dito anteriormente, não existem níveis pré-definidos. Nesse caso, os limites ficam em torno de 93 e 102.

Linha psicológica: porcentagem de dias de subida entre os n dias anteriores. Reflete a força da tendência de subida no período. O indicador varia em um intervalo de 0 a 100, onde valores maiores que 50 indicam tendência de subida e valores menores que 50 tendência de descida.

Fórmula:

$$PSY(n)_t = \left(\frac{UD_n}{n} \times 100\right) \tag{3.20}$$

Exemplo para n = 10:

$$\begin{aligned} PSY(10)_{19} &= \left(\frac{UD_{10}}{10} \times 100\right) \\ PSY(10)_{19} &= \left(\frac{6}{10} \times 100\right) \\ PSY(10)_{19} &= 60 \end{aligned} \tag{3.21}$$

Interpretação do PSY: A Figura 3.13 mostra o gráfico do indicador PSY. Uma linha está desenhada no ponto 50 para separar os valores entre

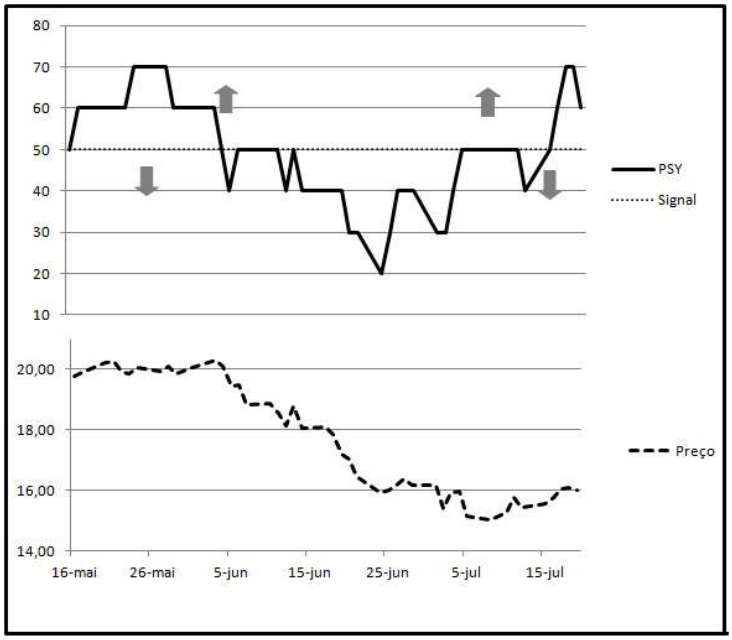

Figura 3.13: Indicador Linha Piscológica e Preço da PETR4 no período de 16 de maio a 19 de julho de 2013

tendência de subida e de descida. Sendo assim, as setas para cima sinalizam a área onde os valores são maiores que 50 e indicam tendência de subida. As setas para baixo, por sua vez, sinalizam a área com tendência de descida.

Disparidade diferença entre o preço atual e a média móvel simples de n dias anteriores. Situa o preço em relação à média do período no qual ele se encontra. Valores do indicador maiores que zero indicam tendência de subida, enquanto valores menores que zero indicam tendência de descida.

Fórmula:

$$Disp(n)_t = \frac{C_t - MA_n}{MA_n} \times 100 \tag{3.22}$$

Exemplo para n = 5:

$$\begin{aligned} Disp(5)_{19} &= \frac{C_{19} - MA_5}{MA_5} \times 100 \\ Disp(5)_{19} &= \frac{15,98 - 15,88}{15,88} \times 100 \\ Disp(5)_{19} &= 0,59 \end{aligned} \tag{3.23}$$

Interpretação Disp: O gráfico da Figura 3.14 mostra os valores do indicador Disp. Note que o indicador manteve-se abaixo de zero grande parte do período. Isso ocorre justamente porque a tendência predominante foi de descida.

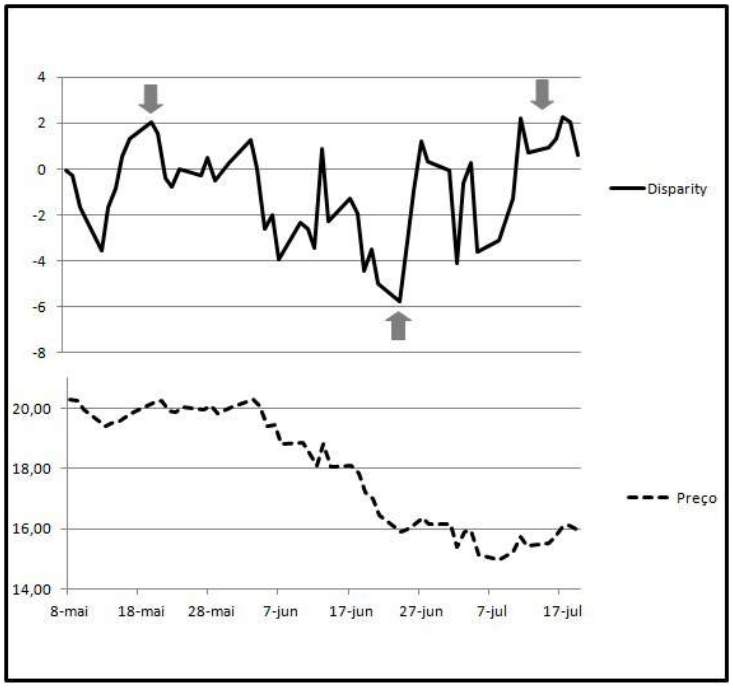

Figura 3.14: Indicador de Disparidade e Preço da PETR4 no período de 8 de maio a 19 de julho de 2013

3.0.9 Medidas de desempenho financeiro

Estratégias de predição de ações, independentes da origem ou da maneira como são realizadas, tem o desempenho avaliado com medidas financeiras específicas. A medida mais essencial é a rentabilidade, obtida pela diferença entre um valor investido x, e o valor y após determinada estratégia de investimento i.

Além da rentabilidade, algumas medidas específicas são sugeridas pela literatura e descritas a seguir:

Retorno sobre investimento (ROI) : quociente entre o lucro e o investimento.

Retorno sobre ativo (ROA) : quociente entre o lucro operacional e o ativo total.

Retorno sobre patrimônio líquido (ROE) : quociente entre lucro líquido e o patrimônio líquido. Mede a rentabilidade dos recursos investidos pelos proprietários.

Lucro por ação (LPA) : quociente entre o lucro líquido e número de ações da empresa

Índice preço/lucro (P/L) : quociente entre o preço de mercado da ação e o lucro por ação.

CAPÍTULO

4

Representação de Classe

Ao representar os dados para algoritmos de aprendizados de máquina, a representação de classe é um dos pontos-chave na abstração de conhecimento do mundo real. Como ilustrado na Figura 4.1, a etapa de representação de classe envolve a rotulação do conjunto de exemplos e gera como saída o conjunto de treino usado na indução do classificador. Assim, o mesmo conjunto de exemplos pode ter a classe representada de maneiras distintas. Consequentemente, os classificadores induzidos com os respectivos conjuntos de treino podem ter desempenho diferente.

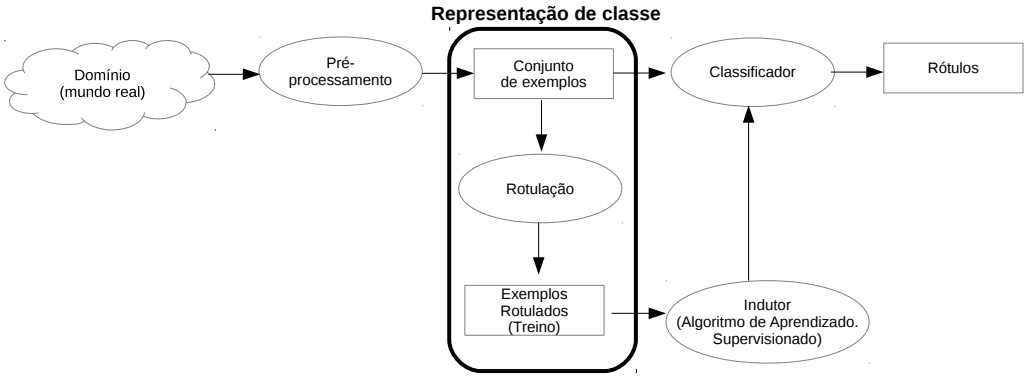

Figura 4.1: A etapa de representação de classe no Aprendizado de Máquina

Nesse sentido, especialmente no problema de predição de ações, a classe tem sido representada com o comportamento de preço da ação. A maioria dos trabalhos utiliza regressão ou tendência de movimento de preço futura, para rotular os exemplos. Entretanto, tal representação pode não gerar o melhor desempenho no ambiente financeiro, por não levar em consideração

a intensidade de variação de preço e se restringir a preços de intervalos de tempo pequenos.

Por outro lado, acredita-se que ao considerar somente as variações mais relevantes, entre preços de um período de análise maior, pode ser mais interessante para representar as classes. Nesse sentido, este trabalho propõe uma representação de classe que identifica pontos de máximo e de mínimo na série temporal de preço da ação.

O restante desta seção descreve os detalhes das representações de classe e o respectivo mapeamento em operações de compra e venda de ações. Na Seção 4.1 é descrita a representação SOBEDESCE, enquanto na Seção 4.2 são apresentados detalhes da utilização de RDP para representar a classe do problema de predição. Por fim, na Seção 4.3 é apresentada a representação de classe LMINMAX proposta.

Por questões de simplicidade, no restante deste trabalho considera-se a observação de preços diários das ações. Entretanto, nada impede que as abordagens propostas sejas implementadas, por exemplo, em preços observados a cada minuto, hora ou semana.

4.1 Sobe Desce

A representação de classe SOBEDESCE (jae Kim, 2003) se baseia na diferença de preço absoluta entre o dia atual (d) e o dia imediatamente posterior ao atual ($d+1$). No valor da classe é puramente refletida a tendência de preço futura da ação, subida ou descida, sem distinguir a intensidade dessa variação.

Assim, o exemplo de um dia qualquer d é rotulado a partir da subtração do preço da ação no dia d, pelo preço da ação no dia $d+1$, sendo a classe:

DESCE: quando a diferença é maior que 0, ou seja, o preço do dia d é maior que o preço do dia $d+1$

SOBE: quando a diferença é menor que 0, ou seja, o preço do dia d é menor que o do dia $d+1$

Um exemplo da aplicação da abordagem SOBEDESCE é ilustrado na Figura 4.2, na qual é mostrada a série temporal de cotações diárias da ação PETR4, entre 18/09/2013 e 17/10/2013. Observe que cada dia do período é rotulado com uma das classes SOBE ou DESCE e que, de fato, a compra e venda de ações nos dias rotulados com sobe e desce, respectivamente, tendem a ser lucrativas.

Entretanto, a quantidade excessiva de operações em um pequeno intervalo de tempo pode ser um ponto fraco da abordagem SOBEDESCE. Isso porque

as corretoras financeiras, autorizadas a operar na bolsa de valores, cobram uma taxa de corretagem por operação, o que pode fazer com que o lucro não compense o custo associado.

Ao contrário da abordagem SOBEDESCE, estratégias com operações diárias como *day trade* são conhecidas por bom desempenho, porque os analistas selecionam os pontos de mínimo e máximo diários, de modo a alcançar o maior lucro possível com apenas duas operações. Enquanto no SOBEDESCE o período observado é maior que um dia, e busca encontrar sequências de operações de compra e venda que, juntas, tendem a gerar algum lucro. Observe que no *day trade* a diferença entre o preço de compra e o de venda é determinante, e é o principal foco dos investidores, porque deseja-se que a diferença compense o custo das duas operações realizadas. Por outro lado, para a representação de classe SOBEDESCE interessa apenas se diferença é positiva ou não, não importando se ela é menor que o custo de corretagem ou se existem outros dias para os quais o lucro seria maior.

Uma solução para tornar a rotulação SOBEDESCE um pouco mais competitiva, no mercado financeiro, seria calcular a variação de preço em um período maior que o entre o dia atual e o próximo. Uma situação que possibilita isso são as sequências de dias com classes iguais por exemplo, nas quais se apenas o primeiro dia é rotulado com SOBE ou DESCE, assim o custo e a quantidade de operações são diminuídos, consequentemente existem mais chances do lucro obtido compensar o custo de corretagem. Uma abordagem nesse sentido é a rotulação com RDP, descrita na Seção 4.2

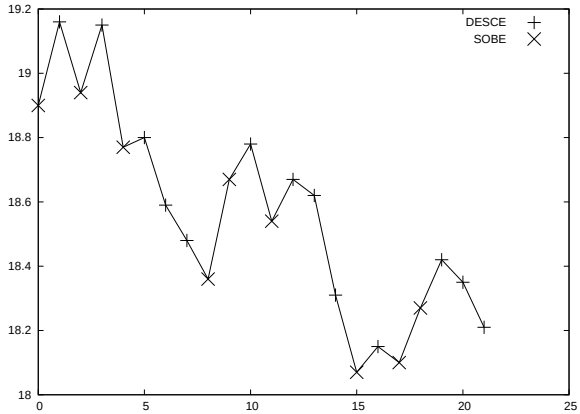

Figura 4.2: Rotulação com Sobe Desce

4.2 RDP

A representação de classe RDP (Cao and Tay, 2003) utiliza a Diferença Relativa Percentual (RDP) entre médias móveis exponenciais. O preço do dia d é transformado em um valor relativo percentual, comparado aos preços do

período em que d se encontra. Ao contrário da representação de classe SOBEDESCE, que utiliza preços pontuais no cálculo da classe, o RDP busca situar o preço do exemplo em relação à média de preços futuros, mais especificamente de 5 dias após o dia d. A equação 4.1 mostra a fórmula aplicada para encontrar o RDP do exemplo do dia d.

$$RDP = \frac{MME(3)_{ct+5} - MME(3)_{ct}}{MME(3)_{ct}} * 100 \tag{4.1}$$

onde ct é o dia atual e MME(3) é a Média Móvel Exponencial com $n = 3$

O resultado do RDP é um valor contínuo que busca representar a variação entre o dia atual e 5 dias futuros. Para convertê-lo em classes discretas utiliza-se uma variação v como limiar. Os exemplos cujo RDP é maior que a variação v positiva são classificados com MAX; aqueles cujo valor de RDP é menor que v negativo são classificados com MIN; e os demais exemplos, com RDP entre $-v$ e $+v$, são classificados com IRR.

4.3 LMINMAX

As abordagens SOBEDESCE e RDP citadas anteriormente são baseadas apenas na tendência futura de preço da ação. Por outro laso, a representação de classe LMINMAX proposta por este trabalho é resultado da observação de um período de tempo maior, que engloba o passado e futuro próximo ao dia que se deseja rotular.

Identificando pares de pontos de mínimo e máximo, para os quais pares de operações de compra e venda de ações são mais lucrativas, a técnica proposta consiste no deslizamento de uma janela sobre a série temporal de preço da ação. A cada deslizamento, os dois pontos relevantes locais da janela são identificados e um voto de máximo e de mínimo é somado para cada valor.

Para ilustrar a técnica, na Figura 4.3 é ilustrado um deslizamento da janela, na série temporal de preço da PETR4 no ano de 2013. Observe que os dias 04/03/2013 (R$ 16.18- 41º dia) e 05/07/2013 (R$ 20.62 - 83º dia) são identificados como o máximo e o mínimo da janela e cada um recebe um voto. Em seguida, na Figura 4.4 são apresentadas três janelas $w1$, $w2$ e $w3$ de tamanho $k = 75$, sobre a mesma série temporal. Observe, nesse caso, que o dia 41 é mínimo nas três janelas, contabilizando assim três votos para o dia 41 como dia de mínimo. O mesmo acontece com o dia 83, sendo máximo nas três janelas, contabilizando três votos para máximo.

Depois que a janela deslizante percorre todo o período, a quantidade de votos dos pontos relevantes de máximo e de mínimo é ordenada de maneira decrescente. Os primeiros r dias com os maiores somatórios de votos são rotulados como MAX ou MIN. Em seguida, no sentido de aumentar a quantidade

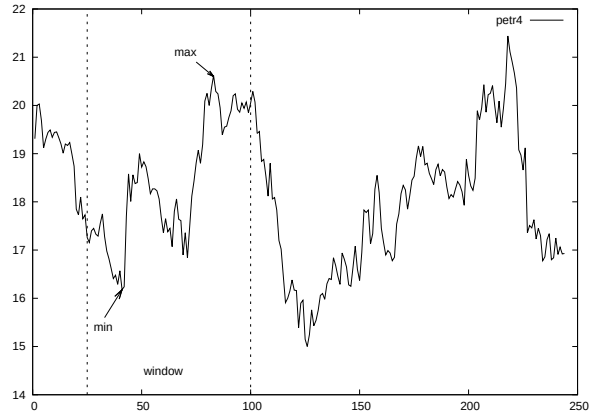

Figura 4.3: Definição da janela $W^*_{PETR4} = (25, 100)$ do intervalo W_{PETR4}

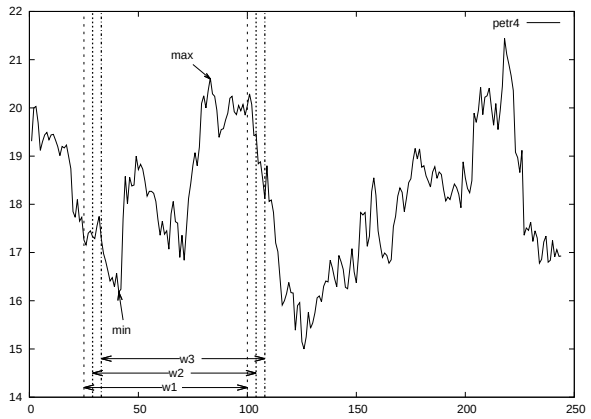

Figura 4.4: Janela deslizante para encontrar os pontos de máximo e mínimo locais. Dias 41 e 83 são o máximo e mínimo local das três janelas

de pontos relevantes, dos dias imediatamente anteriores aos rotulados como máximos e mínimos são selecionados aqueles cujo preço tende a ser máximo e mínimo, observando uma variação v. Por fim, todos os demais dias são rotulados como IRR.

Nesse sentido, considerando cada dia da série temporal como um exemplo da representação de dados, a técnica LMINMAX proposta consiste em uma função de rotulação dos exemplos, de modo a representar a classe em três valores possíveis: MAX, MIN e IRR.

4.3.1 Notação

Como descrito na seção anterior, a proposta deste trabalho consiste em definir uma função de rotulação l, que resulta em pontos relevantes (máximo e mínimo) e irrelevantes. A função proposta identifica os pontos relevantes locais e contabiliza a frequência no período como um todo, rotulando os exemplos de acordo com a quantidade de vezes que o preço do exemplo é tido como relevante no período. Observe, a seguir, a definição da identificação dos pontos de máximo e mínimo locais.

Definição 1 (w^*: máximo e mínimo local de um W)
Considere um intervalo de tempo entre $data_{ini}$ e $data_{fim}$ e p_i o preço da ação do i-ésimo dia de operação da bolsa de valores durante esse intervalo de tempo. Uma janela de tempo W_d é um intervalo de dias definido por $W_d = [data_{ini}, data_{fim}$ e $W = [p_1, ..., p_n]$ é o vetor de preços da ação, onde n é a quantidade de dias de abertura da bolsa no período W_d, ou seja, ignorando os finais de semana, feriados e dias em que a bolsa não operou. A janela de busca local w é uma janela contida em W.

Sejam $ipreco\_min(w)$ *e* $ipreco\_max(w)$ *os índices dos valores mínimo e máximo de* w*. Quando* $ipreco\_min(w) \prec ipreco\_max(w)$*, ou seja, quando a data* $ipreco\_min(w)$ *precede a data* $iprice\_max(w)$*, o mínimo e máximo local são dados pela tupla* $w^* = (ipreco\_min(w), ipreco\_max(w))$*.*

Para exemplificar a definição, considere a série temporal de preço da PETR4 de 2013, ilustrada na Figura 4.5 que consiste na janela $W_d = [02/01/2013, 31/12/2013]$. Observe que, ao limitar o investidor a apenas uma compra e uma venda durante o período da figura, o maior lucro é obtido pela compra da ação no dia 126 (08/Julho), a R$ 15,00, e venda no dia 218 (18/Novembro) a R$ 21,44. Essas duas datas compõem os pontos de mínimo e máximo de W_{PETR4}, representados por $W^*_{PETR4} = (ipreco\_min(W_{PETR4}), ipreco\_max(W_{PETR4})) = (126, 218)$ e cujo lucro é representado por $L(W^*_{PETR4}) = W_{PETR4}[ipreco\_max(W_{PETR4}] - W_{PETR4}[ipreco_min(W_{PETR4})] = 6,44$.

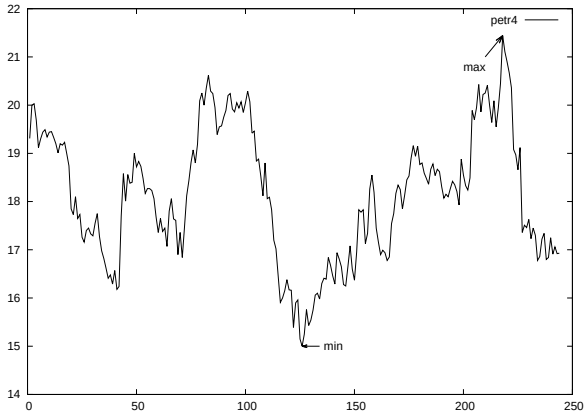

Figura 4.5: Máximo e mínimo de $W_{PETR4} = (126, 218)$

Pela restrição de apenas uma compra e uma venda em um determinado período (janela) W, sabe-se que não existem outros dois pontos (dias) que possam oferecer ao investidor um lucro maior $L(W^*_{PETR4}) = R\$6,44$ para PETR4 na janela W_{PETR4}. Por exemplo, ao analisar os intervalos de tempo $w \subseteq W_{PETR4}$, como a janela entre 06/02/2013 (25^o dia) e 31/05/2013 (100^o dia) é possível encontrar outros pontos de mínimo (R$ 16.18 - 03/04/2013 - 41^o dia) e máximo (R$ 20.60 - 07/05/2013 - 83^o dia) que representam um lucro de R$ 4,42. Note que, nessa situação, outros pontos máximos e mínimos possibilitam lucro igual ou inferior que $W*$.

Na Figura 4.3 são mostrados os intervalos de máximo e mínimo diferentes encontrados no intervalo W_{PETR4} e que sempre terão lucro menor ou igual ao lucro do W_{PETR4}. Existem outros períodos onde os pontos de máximo e mínimo serão iguais a W_{PETR4}, basta que $w \subseteq W^*_{PETR4}$.

Quando $w \subset W$, ou seja, intervalos menores w contidos em W são definidos, é possível encontrar mais intervalos w^* que podem ser diferentes de

W^*. Assim, com diferentes intervalos w^* é possível fazer diversas operações de compra e venda, onde a soma dos lucros desses w^* é maior que W^*.

4.3.2 Algoritmo proposto

A proposta deste trabalho consiste em encontrar os pontos de mínimo e máximo indicados por todos os w^* de $w \subset W$ onde o tamanho da janela $k = |w|$ é fixo. Para isso as janelas de tamanho $|w|$ contidas em W são encontradas por meio de uma janela deslizante de tamanho k sobre a série de preços W. Assim a janela w é deslocada em um dia, ou seja, em uma estrutura de fila é adicionado um novo dia e removido o último dia da janela em cada iteração do algoritmo. Para cada janela w são calculados o $ipreco\_max(w)$ and $ipreco\_min(w)$. Cada vez que um dia é indicado como ponto de máximo, o método "vota" no dia como máximo. Cada vez que um dia é indicado como ponto de mínimo, o método "vota" no dia como mínimo. Os dias com maiores quantidade de votos são rotulados como máximo (MAX) e mínimo (MIN) e os demais dias são, inicialmente, rotulados como irrelevantes (IRR).

Desse modo, somente preços mínimos e máximos pontuais são rotulados, resultando em um conjunto de dados bastante desbalanceado, porque a maioria dos exemplos é IRR. Para incrementar as classes minoritárias, assume-se que os dados imediatamente anteriores aos mínimos e máximos possuem informação relevante e similar ao MIN e MAX. Desse modo, o método observa 10% da janela w imediatamente anterior aos dias de máximo e mínimo. Os preços desses dias são convertidos para *scores* z ($z = \frac{w(d)-\mu}{\sigma}$). Os *scores* z que tem diferença menor que 20% na CDF (*Cumulative Distribution Function*) do dia rotulado, são rotulados com a mesma classe. O Algoritmo 1 ilustra o pseudo código para encontrar os pontos de máximo e de mínimo utilizando a metodologia descrita e na Figura 4.6 é ilustrado um exemplo de rotulação com a representação de classe LMINMAX.

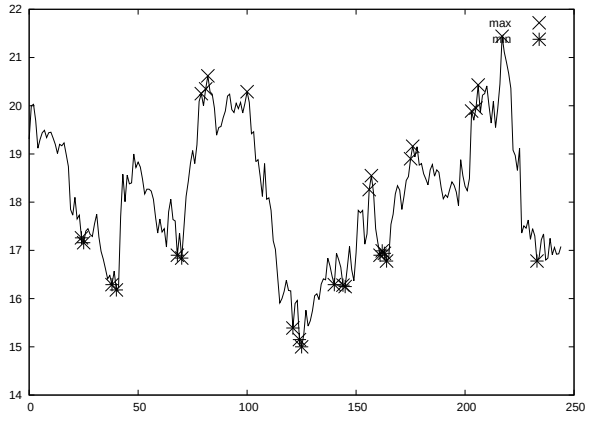

Figura 4.6: Dias rotulados como MAX e MIN, os dias não marcados são considerados IRR

Algorithm 1: Rotula dias com MAX, MIN,IRR

Input: W: vetor de dias com os valores da ação;
k: tamanho da janela de observação;
r: limite do ranking de contagem de máximos e mínimos;
Output: $vclass$: vetor de dias com as classes MIN,MAX,IRR

$w_{ini} \leftarrow 1$;
$w_{fim} \leftarrow k$;
$w = [w_{ini}, w_{fim}]$;
$v_{max} \leftarrow$ inicializa com zeros;
$v_{min} \leftarrow$ inicializa com zeros;
$vclass \leftarrow$ inicializa com IRR;
while w_{fim} não chegou ao fim de W **do**
    $v_{max}[dmax(w)] \leftarrow v_{max}[dmax(w)] + 1$;
    $v_{min}[dmin(w)] \leftarrow v_{min}[dmin(w)] + 1$;
    desloca janela w em um dia;
end
$ordmax \leftarrow$ os maiores r valores v_{max} em ordem decrescente;
$ordmin \leftarrow$ os menores r valores v_{min} em ordem crescente;
foreach dia de $ordmax$ **do**
    $w = [dia - k \times 0.1, dia]$;
    $cdf \leftarrow$ converte w para o CDF do z-score;
    atribui classe $vende$ para os dias $i \in w$ que possuem $cdf[i] > (cdf[dia] - 0.2)$ em $vclass$
end
foreach dia de $ordmin$ **do**
    $w = [dia - k \times 0.1, dia]$;
    $cdf \leftarrow$ converte w para o CDF do z-score;
    atribui classe $compra$ para os dias $i \in w$ que possuem $cdf[i] < (cdf[dia] + 0.2)$ em $vclass$
end
return $vclass$

CAPÍTULO

5

Avaliação Experimental

A avaliação experimental descrita neste capítulo busca comparar o desempenho da representação de classe proposta por este trabalho (LMINMAX), com outras duas representações de classe (SOBEDESCE e RDP) citadas na literatura e descritas nas Seções 4.1 e 4.3.

Os experimentos foram divididos em seis etapas, ilustradas na Figura 5.1 e descrita a seguir:

1. **Coleta:** as cotações são coletadas e dispostas em séries temporais financeiras por ação

2. **Divisão:** as cotações das séries temporais são divididas em treino e teste

3. **Construção:** os conjuntos de dados de treino e teste são construídos por ação e representação de classe

4. **Indução:** os conjuntos de treino, juntamente com os algoritmos de aprendizado, são utilizados para induzir classificadores

5. **Classificação:** os exemplos do conjunto de teste são apresentados aos respectivos classificadores para predição

6. **Simulação:** são simuladas operações de compra e venda de ações a partir das classes atribuídas aos exemplos de teste.

As Seções 5.1, 5.3, 5.2, 5.4, 5.5 e 5.6 descrevem cada uma das etapas, respectivamente.

A

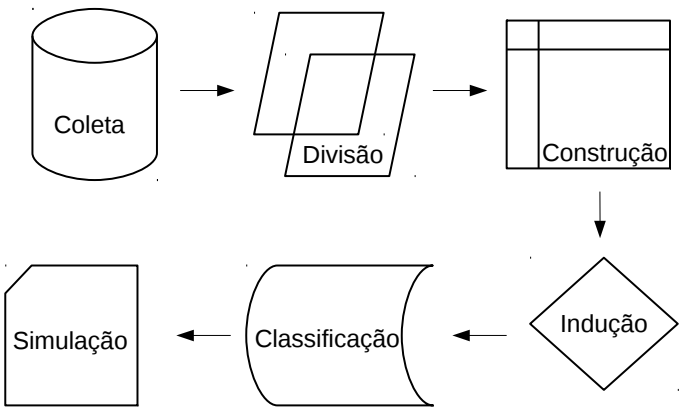

Figura 5.1: As seis etapas da avaliação experimental: coleta, divisão, construção, indução, classificação e simulação

5.1 Coleta dos Dados

Os dados da avaliação experimental deste trabalho foram obtidos Bolsa de Valores de São Paulo (BOVESPA). As ações do conjunto experimental foram selecionadas considerando a respectiva participação no índice BOVESPA (IBOVESPA), citado na Seção 3.0.6. Considerando que o IBOVESPA é composto pelas ações com maior negociabilidade e representatividade do mercado de ações brasileiro, as cerca de 70 ações que o compõem tem participação entre aproximadamente 0.01% e 10% nas negociações. Por isso foram selecionadas nove ações, com participação entre 2% e 10%, para compor o conjunto experimental. Na Tabela 5.1 são listadas as ações selecionadas, com informações retiradas do site da BOVESPA em 29/08/2013. A primeira coluna da tabela mostra a sigla da ação, a segunda coluna o nome da empresa da ação e, a terceira, a participação percentual da ação no índice IBOVESPA. Assim, o conjunto experimental foi composto pelas ações: ABEV3, BBAS3, BBDC4, BRFS3, ITSA4, ITUB4, PETR3, PETR4, VALE3 e VALE5.

Ação	Empresa	Part. no IBOVESPA (%)
ABEV3	AMBEV S/A	5.22
BBAS3	BRASIL	2.8
BBDC4	BRADESCO	7.83
BRFS3	BR FOODS	3.57
ITSA4	ITAU SA	3.16
ITUB4	ITAU UNIBANCO	1.99
PETR3	PETROBRÁS	5.62
PETR4	PETROBRÁS	8.9
VALE3	VALE3	3.58
VALE5	VALE5	4.45

Tabela 5.1: Participação no índice IBOVESPA das ações do conjunto experimental

As informações históricas financeiras foram coletadas para os anos entre

2008 e 2013, do portal *Yahoo Finance*, que fornece um histórico completo de mais de dez anos das ações da BOVESPA. O método *quotes_historical_yahoo* da biblioteca *matplotlib finance* é disponibilizado para extração das cotações automaticamente. Utilizando esse método com o parâmetro $adjusted = False$, os preços não ajustados de abertura, fechamento, máximo e mínimo, além do volume negociado, foram coletados para cada dia do período. O ajuste de preços não foi considerado no contexto deste trabalho porque não é a representação de dados proposta ainda não distingue as variações de preço ocasionadas naturalmente no decorrer do tempo, daquelas resultantes de operações manuais pontuais, tais como desdobramento (*split*).

Os dados coletados totalizaram em torno de 1500 dias de operação da bolsa para cada ação, o que possibilitou aplicação da abordagem proposta em períodos de tempo com características distintas.

5.2 Divisão em Treino e Teste

Após a coleta, os dados brutos utilizados na construção dos conjuntos de dados são divididos em treino e teste, para garantir que os conjuntos de dados posteriormente construídos sejam disjuntos.

Como rotina de divisão utiliza-se *walk-forward* (Kaastra and Boyd, 1996), na qual os dados em ordem cronológica constituem o conjunto de treino-teste, ilustrado na Figura 5.2, de maneira que o conjunto de teste seja composto pelos dias imediatamente posteriores ao treino, e nunca o contrário.

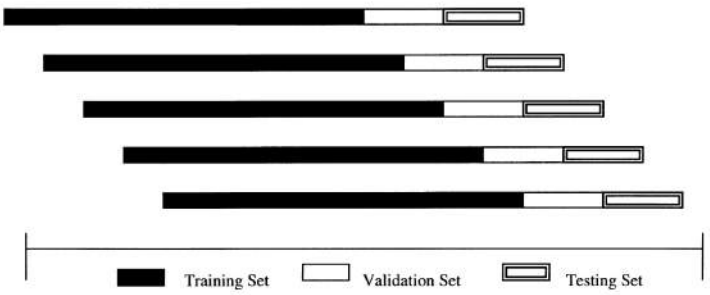

Figura 5.2: Divisão da série temporal em subconjuntos de treino e teste (Kaastra and Boyd, 1996)

Ao respeitar a ordem cronológica, os dados da intersecção entre o treino e teste são desconsiderados em ambos os conjuntos, garantindo a disjunção. Isso é necessário porque o cálculo dos indicadores econômicos (do teste) utiliza dados passados (do treino); e o cálculo da classe (do treino) utiliza dados futuros (do teste).

A fim de avaliar a classificação sob mais de uma perspectiva, os dados foram divididos em diferentes períodos (*splits*), entre os anos de 2008 e 2013.

Variando a quantidade de meses, os meses e anos de início e fim, e utilizando aproximadamente dois anos para treino e um para teste, as seguintes divisões (*splits*) foram realizadas:

01/2008 - 12/2009 - 12/2010: treino de janeiro de 2008 a dezembro de 2009 e teste de janeiro de 2010 a dezembro de 2010.

05/2008 - 04/2010 - 04/2011: treino de maio de 2008 a abril de 2010 e teste de maio de 2010 a abril de 2011.

01/2009 - 12/2010 - 12/2011: treino de janeiro de 2009 a dezembro de 2011 e teste de janeiro de 2011 a dezembro de 2012.

08/2009 - 07/2011 - 07/2012: treino de agosto de 2009 a julho de 2011 e teste de agosto de 2011 a julho de 2012.

01/2010 - 12/2011 - 12/2012: treino de janeiro de 2010 a dezembro de 2012 e teste de janeiro de 2012 a dezembro de 2013.

10/2010 - 09/2012 - 09/2013: treino de outubro de 2010 setembro de 2012 e teste de outubro de 2012 a setembro de 2013.

01/2011 - 12/2012 - 12/2013: treino de janeiro de 2011 a dezembro de 2012 e teste de janeiro de 2013 a dezembro de 2013.

5.3 Construção dos Conjuntos de Dados

A construção dos conjuntos de dados é específica por ação. Os exemplos que compõem cada conjunto mapeiam o comportamento diário da ação na BOVESPA por meio de uma representação de atributos composta indicadores econômicos. Foram considerados os indicadores citados na literatura em (jae Kim, 2003), (Huang and Tsai, 2009), (Klassen, 2005), além da cotação da ação no dia do exemplo.

O cálculo de cada indicador é descrito na Seção 3.0.8 e considera o dia atual como o dia do exemplo. Uma lista com todos os atributos que compõem os exemplos pode ser visualizada na Tabela 5.2.

O atributo classe, por sua vez, foi calculado sobre a série temporal de preço de fechamento da ação, para cada representação de classe (SOBEDESCE, RDP,LMINMAX). A combinação das três representações de classe, com as nove ações do conjunto experimental, implicou a construção de 36 conjuntos de dados para cada período considerado. Para exemplificar, na Tabela 5.3 é ilustrada a distribuição de exemplos por classe para um conjunto de dados de cada ação.

Nº atributo	Sigla/Nome	Descrição
1	Fechamento	Preço de fechamento do dia atual
5	MMS(10)	Média Móvel Simples de 10 dias
6	MME(5)	Média Móvel Exponencial de 5 dias
7	MACD	Convergência/Divergência de Média Móvel
8	IRF(5)	Índice de Força Relativa de 5 dias
9	IRF(14)	Índice de Força Relativa de 14 dias
10	K(9)	Estocástico K de 9 dias
11	K(14)	Estocástico K de 14 dias
12	LARRY(10)	Indicador Larry Williams de 10 dias
13	KD(9)	Estocástico D de 9 dias
14	KD(14)	Estocástico D de 14 dias
15	MOM(3)	Indicador de momento de 3 dias
16	MOM(5)	Indicador de momento de 5 dias
17	MUD(3)	Indicador de mudança de 3 dias
18	MUD(5)	Indicador de mudança de 5 dias
19	PSY(10)	Linha Psicológica de 10 dias
20	DISP(5)	Disparidade de 5 dias
21	BBSUP(5)	Bollinger Bands Superior
22	BBINF(5)	Bollinger Bands Inferior
23	Classe	Atributo classe

Tabela 5.2: Representação de atributos dos conjuntos de dados

Ação	LMinMax	RDP	SobeDesce
ABEV3	(36,661,34)	(81,439,201)	(353,418)
BBAS3	(36,660,33)	(160,357,202)	(358,411)
BBDC4	(34,579,33)	(138,334,164)	(337,349)
BRFS3	(33,660,37)	(151,384,185)	(340,430)
ITSA4	(34,639,35)	(147,354,197)	(355,393)
ITUB4	(40,624,34)	(151,351,186)	(359,379)
PETR3	(31,660,40)	(197,343,181)	(377,394)
PETR4	(34,617,40)	(185,326,170)	(363,368)
VALE3	(36,657,38)	(168,375,178)	(371,400)
VALE5	(34,619,39)	(147,373,162)	(374,358)

Tabela 5.3: Exemplo da distribuição desbalanceada de classes do problema de predição de ações (MAX,IRR,MIN)

5.4 Indução dos Classificadores

Na indução de classificadores foram testados dois algoritmos, SVM e KNN, ambos descritos nas Seções 2.2.1 e 2.2.2.

O algoritmo SVM foi considerado por se tratar do estado da arte, em diversas tarefas de classificação, e por conta da expressiva quantidade de trabalhos na área de predição de ações que o sugerem como o algoritmo mais promissor (Huang and Tsai, 2009). Já o KNN, foi escolhido porque apesar da simplicidade tem se mostrado relevante em problemas de séries temporais, superando os algoritmos conhecidos como árvores de decisão, redes neurais, redes *Bayesianas* e SVM, como (Xi et al., 2006) mostra em sua revisão bibliográfica.

A indução dos classificadores foi realizada com o auxílio da biblioteca *Scikit-learn* do *Python* e com o ajuste de parâmetros automático proposto por (Bergstra and Bengio, 2012). Buscando otimizar a medida de AUC, descrita na Seção 2.2.4, o ajuste resultou em valores ótimos para os parâmetros c e $gamma$ do SVM e número de vizinhos mais próximos do KNN.

5.5 Classificação

Com os classificadores induzidos na Seção 5.4, os limiares de decisão foram calibrados otimizando a medida de AUC, por causa das classes desbalanceadas. Para tanto, os valores de confiança cujos pontos na curva ROC são os mais próximos ao ponto (0,1) foram selecionados como limiares. Para exemplificar, considere os valores de confiança ordenados por *rank* da Tabela 5.4 e a curva ROC da Figura 5.3. Observe que o ponto 0.67 ilustrado na figura corresponde ao oitavo exemplo do *rank* da tabela, e representa o ponto mais próximo ao canto superior esquerdo. Assim, ele é o valor selecionado como limiar de decisão do classificador.

Exemplo	Pontuação (valor de confiança)	Classe verdadeira	Direção na curva ROC
1	0.99	positiva	cima
2	0.91	positiva	cima
3	0.9	positiva	cima
4	0.88	positiva	cima
5	0.8	positiva	cima
6	0.77	positiva	cima
7	0.75	negativa	direita
8	0.67	positiva	cima
9	0.67	positiva	cima
10	0.62	negativa	direita
11	0.59	positiva	cima
12	0.51	negativa	direita
13	0.5	negativa	direita
14	0.45	negativa	direita
15	0.4	negativa	direita
16	0.38	negativa	direita
17	0.3	negativa	direita
18	0.2	negativa	direita
19	0.18	negativa	direita
20	0.11	negativa	direita

Tabela 5.4: Ordenação por *rank* dos valores de confiança atribuídos por um classificador a um conjunto de exemplos rotulados e as respectivas direções na curva ROC resultantes das combinações de ordenação e classe verdadeira

Aos classificadores induzidos e calibrados são apresentados os exemplos de teste. Os rótulos preditos pelo classificador para os exemplos de teste

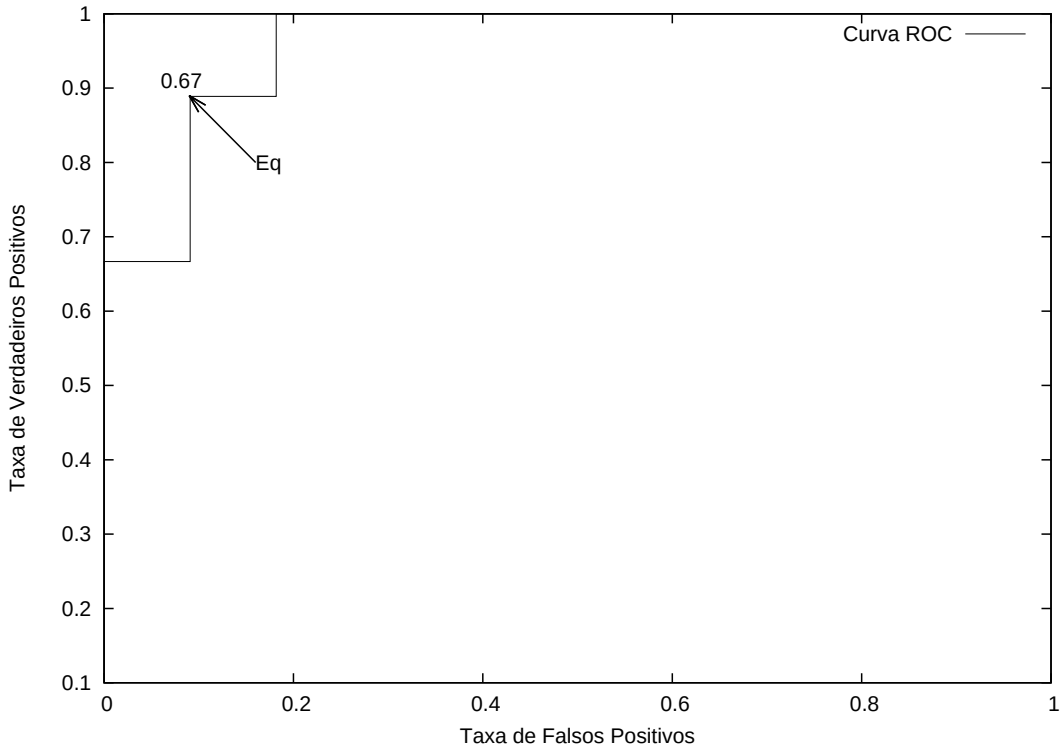

Figura 5.3: Curva ROC resultante da ordenação por *rank* da Tabela 5.4, com o limiar de decisão que otimiza a AUC identificado em 0.67

são comparados com os rótulos verdadeiros (esperados) e, os resultados de classificação são contabilizados com as medidas descritas na Seção 2.2.4.

Conforme a necessidade, os limiares de decisão dos classificadores são calibrados de maneira a otimizar a acurácia de classificação.

Por fim, os rótulos preditos são fornecidos como entrada para a etapa de simulação, descrita na Seção 5.6.

5.6 Simulação

A saída gerada pela classificação, descrita na Seção 5.5, é um conjunto de classes atribuídas aos exemplos de teste. Na simulação essas classes são interpretadas como operações de compra e venda de ações, seguindo estratégias condizentes com a representação de classe do conjunto de dados. Assim, são definidas três estratégias de simulação, uma para cada representação de classe.

Para as representações SOBEDESCE e RDP a simulação realiza uma operação de compra toda vez que a classe predita é SOBE, e as vendas são realizadas em momentos distintos. No LMINMAX as compras e vendas são sinalizadas pelo classificador. Assim, as três estratégias são as seguintes:

SOBEDESCE calcula a variação diária imediatamente futura, então a venda é realizada um dia após a compra.

RDP utiliza 5 dias futuros para determinação da classe, assim o simulador vende cinco dias após a compra

LMINMAX além de compra o simulador também vende de acordo com a classe predita pelo classificador

Além da estratégia de simulação, outro ponto relevante é o preço utilizado como referência para simular a operação. Como o preço de fechamento do dia é um dos atributos do exemplo, a predição do classificador só é conhecida após o encerramento diário da bolsa. Consequentemente, as operações precisam ser realizadas com o preço do seguinte à predição.

O processo de simulação das operações de um intervalo de tempo W, parte de um investimento inicial i, um saldo s (que inicialmente recebe o investimento i) e uma quantidade de ações q (iniciada em 0). Nas operações de compra, o saldo s é convertido em ações, igualado a 0 e a quantidade de ações q é atualizada para o valor do saldo dividido pelo preço da ação p. Por outro lado, nas operações de venda a quantidade de ações q é convertida em dinheiro, igualada a 0 e o saldo s recebe o produto da quantidade de ações q pelo preço da ação p. Além disso, a fim de tornar a simulação mais próxima do ambiente real de investimento, considera-se a taxa de corretagem das operações. O valor foi estabelecido em R$10 por operação, uma vez que os valores praticados para a IBOVESPA variam em torno de R$2.9 e R$25.21, dependendo da corretora e do tipo de contrato com o investidor. Nas fórmulas 5.1 e 5.2 são apresentados os cálculos realizados a cada operação de compra e venda, respectivamente.

Operação de compra:

$$\begin{aligned} saldo &\leftarrow saldo - taxa_{corretagem} \\ qtde_{acoes} &\leftarrow \frac{saldo}{preco_{acao}} \\ saldo &\leftarrow 0 \end{aligned} \tag{5.1}$$

Operação de venda:

$$\begin{aligned} saldo &\leftarrow qtde_{acoes} \times preco_{acao} \\ saldo &\leftarrow saldo - taxa_{corretagem} \\ qtde_{acoes} &\leftarrow 0 \end{aligned} \tag{5.2}$$

onde s é o saldo, c a taxa de corretagem e q a quantidade de ações

Como pode-se ver, a cada operação todas as ações são vendidas ou todo o saldo é comprado em ações, o que impossibilita a realização de duas operações

iguais subsequentes. Tal característica é relevante porque deseja-se que, em situações nas quais tem-se uma série de operações iguais, apenas a primeira seja realizada para evitar altos custos de corretagem.

Ao final da simulação das operações de um intervalo W, a rentabilidade é calculada com a diferença percentual entre o saldo final e o investimento inicial i, conforme descrito na Equação 5.3. Neste trabalho o intervalo W para o qual foi calculada a rentabilidade foi anual e de três anos (2011, 2012, 2013).

$$Rentabilidade = (\frac{saldo_{final}}{investimento_{inicial}} - 1) * 100 \tag{5.3}$$

O Exemplo 1 a seguir mostra a simulação de uma sequência de operações de compra e venda fictícias. Exemplo 1: Considere os preços listados na Tabela 5.5, um investimento inicial de R$10000 e uma taxa de corretagem de R$5 por operação. A primeira operação é uma compra, sinalizada na linha 1, no dia 11/11. Então a compra é realizada no dia 12/11, com o preço de abertura de R$20.06. O saldo é atualizado com a dedução da taxa de corretagem, dividido pelo preço, resultando em 498.25 ações, e zerado. No dia 12/11 nenhuma operação é sinalizada e, no dia 13/11, uma operação de compra é sinalizada mas não pode ser realizada porque não há saldo disponível. Na linha 4, dia 14/11, é sinalizada uma operação de venda, que é realizada na abertura do dia seguinte, 18/11, com o preço de R$20. O saldo, então, é atualizado para o produto da quantidade de ações pelo preço da ação e depois recebe a dedução da taxa de corretagem, totalizando R$10209.23. Observa-se uma rentabilidade de R$209.23. Além disso a quantidade de ações é zerada. Na linha 5 nenhuma operação é sinalizada e, na linha 6, uma operação de venda é sinalizada mas não pode ser realizada porque não há ação disponível para venda. Assim a simulação chega ao fim com um lucro de aproximadamente 2%. Um detalhe quanto à finalização da simulação é que, se existem ações para serem vendidas no último dia, é realizada uma venda compulsória independente da classe do exemplo do último dia.

Linha	Dia	Preço	Sinal	Saldo (R$)	Ações
1	11/11/2013	19.66	Compra	10000	0
2	12/11/2013	20.06	-	0	498.25
3	13/11/2013	19.55	Compra	0	498.25
4	14/11/2013	20	Vende	0	498.25
5	18/11/2013	20.5	-	10209.23	0
6	19/11/2013	21.3	Vende	10209.23	0
7	21/11/2013	20.5	Compra	10209.23	0
8	22/11/2013	20.75	-	0	491.77
9	25/11/2013	20.83	-	0	491.77

Tabela 5.5: Informações das operações de compra e venda simuladas no Exemplo 3

5.7 Resultados

Nesta seção são descritos os resultados da avaliação experimental do Capítulo 5, organizados em questões de pesquisa. Elaboradas durante a avaliação experimental, buscando refinar e ajustar os parâmetros dos algoritmos utilizados, cada questão implica na seleção de uma técnica, algoritmo ou valor de parâmetro. A cada resposta, considera-se a técnica, o algoritmo ou o parâmetro como selecionado para os resultados da questão seguinte.

1. Qual o rendimento na simulação esperado (limite superior) para as abordagens de rotulação RDP, SobeDesce e LMinMAx nos conjuntos de dados do conjunto experimental?

   Na simulação com os rótulos verdadeiros, ou seja, considerando as medidas de desempenho em seus valores máximos, pode-se verificar o lucro esperado para cada conjunto de dados. Nas Tabelas de 5.6 até 5.12 são ilustrados os lucros que seriam obtidos, em cada conjunto de dados, utilizando os três métodos de rotulação comparados.

   Observe que o método de rotulação SobeDesce supera os demais em praticamente todas as simulações, com ganhos esperados de até 328.11%, no *split* de agosto de 2011 até julho de 2012, para a ação PETR3. Em segundo lugar tem-se o método LMinMAx, cujo desempenho é superior que o RDP na maioria dos conjuntos de dados, com ganhos esperados de até 206.91%, no *split* de outubro de 2012 até setembro de 2013, para a ação da PETR3. Por fim, o máximo de ganho esperado para o RDP é de 136.39%, no *split* de agosto de 2011 até julho de 2012, para a ação ITUB4.

Janeiro/2010 até Dezembro/2011			
Ação	LMinMAx (%)	RDP (%)	SobeDesce (%)
ABEV3	129.03	128.36	177.26
BBAS3	146.34	73.82	188.16
BBDC4	183.75	121.50	210.61
BRFS3	109.77	80.73	190.28
ITSA4	125.16	89.56	203.59
ITUB4	162.40	80.57	185.47
PETR3	68.79	72.25	134.02
PETR4	121.18	81.67	146.59
VALE3	147.66	80.21	199.61
VALE5	93.13	83.36	168.95

Tabela 5.6: Rendimento esperado por ação (limite superior) para as estratégias de rotulação RDP,SobeDesce e LMinMAx no conjunto de teste de Janeiro de 2010 até Dezembro de 2010

   Na Figura 5.4 é ilustrado um gráfico da média de rendimento e erro padrão em todos os conjuntos de dados, por método de rotulação. Observe

Maio/2010 até Abril/2011			
Ação	LMinMAx (%)	RDP (%)	SobeDesce (%)
ABEV3	86.31	107.43	183.48
BBAS3	127.96	70.70	179.33
BBDC4	129.34	119.64	206.54
BRFS3	131.66	83.67	202.13
ITSA4	141.00	79.64	178.54
ITUB4	124.89	75.95	172.77
PETR3	125.73	76.45	166.87
PETR4	82.63	62.67	145.86
VALE3	113.11	79.28	209.90
VALE5	81.21	78.64	162.23

Tabela 5.7: Rendimento esperado por ação (limite superior) para as estratégias de rotulação RDP,SobeDesce e LMinMAx no conjunto de teste de maio de 2010 até abril de 2011

Janeiro/2011 até Dezembro/2011			
Ação	LMinMAx (%)	RDP (%)	SobeDesce (%)
ABEV3	147.29	76.33	191.23
BBAS3	151.82	93.28	162.84
BBDC4	108.83	85.67	139.00
BRFS3	200.50	104.18	205.69
ITSA4	157.15	112.23	192.98
ITUB4	152.00	103.01	171.75
PETR3	112.17	82.16	234.68
PETR4	90.79	49.24	124.62
VALE3	98.99	87.56	139.77
VALE5	77.04	68.94	112.89

Tabela 5.8: Rendimento esperado por ação (limite superior) para as estratégias de rotulação RDP,SobeDesce e LMinMAx no conjunto de teste de janeiro de 2011 até dezembro de 2011

Agosto/2011 até Julho/2012			
Ação	LMinMAx (%)	RDP (%)	SobeDesce (%)
ABEV3	137.21	77.68	178.51
BBAS3	135.65	117.92	248.71
BBDC4	95.70	85.16	164.19
BRFS3	95.35	67.94	205.38
ITSA4	160.06	113.79	283.09
ITUB4	139.87	136.39	214.96
PETR3	156.27	89.13	328.11
PETR4	155.72	90.28	207.58
VALE3	106.18	89.74	206.64
VALE5	110.43	83.62	164.14

Tabela 5.9: Rendimento esperado por ação (limite superior) para as estratégias de rotulação RDP,SobeDesce e LMinMAx no conjunto de teste de agosto de 2011 até julho de 2012

que, dentre as três estratégias, a SobeDesce é a que possui maior rendimento esperado; seguida pela abordagem proposta LMinMAx e, por último, RDP.

2. Qual combinação de algoritmo indutor e método de rotulação, entre SVM

Janeiro/2012 até Dezembro/2012			
Ação	LMinMAx (%)	RDP (%)	SobeDesce (%)
ABEV3	187.06	118.34	206.85
BBAS3	167.26	133.28	251.25
BBDC4	95.64	100.70	115.84
BRFS3	154.89	80.31	184.08
ITSA4	126.17	111.44	195.77
ITUB4	68.99	126.57	158.24
PETR3	131.94	82.71	215.54
PETR4	105.44	69.98	168.90
VALE3	175.54	65.40	210.10
VALE5	153.27	69.87	156.71

Tabela 5.10: Rendimento esperado por ação (limite superior) para as estratégias de rotulação RDP,SobeDesce e LMinMAx no conjunto de teste de janeiro de 2012 até dezembro de 2012

Outubro/2012 até Setembro/2013			
Ação	LMinMAx (%)	RDP (%)	SobeDesce (%)
ABEV3	125.74	85.77	216.67
BBAS3	173.72	131.34	214.55
BBDC4	137.33	113.13	190.58
BRFS3	129.27	81.19	153.02
ITSA4	69.55	69.83	153.15
ITUB4	94.25	89.54	209.65
PETR3	206.91	71.21	189.35
PETR4	133.34	52.62	171.45
VALE3	163.46	59.17	259.67
VALE5	145.03	57.38	237.01

Tabela 5.11: Rendimento esperado por ação (limite superior) para as estratégias de rotulação RDP,SobeDesce e LMinMAx no conjunto de teste de outubro de 2012 até setembro de 2013

Janeiro/2013 até Dezembro/2013			
Ação	LMinMAx (%)	RDP (%)	SobeDesce (%)
ABEV3	76.75	44.78	156.18
BBAS3	145.16	99.34	183.77
BBDC4	126.47	87.92	191.71
BRFS3	93.40	50.04	129.90
ITSA4	60.49	59.70	143.89
ITUB4	119.74	80.38	195.12
PETR3	198.51	65.47	218.70
PETR4	124.32	64.98	198.53
VALE3	191.46	60.48	229.45
VALE5	127.78	60.07	207.82

Tabela 5.12: Rendimento esperado por ação (limite superior) para as estratégias de rotulação RDP,SobeDesce e LMinMAx no conjunto de teste de janeiro de 2013 até dezembro de 2013

e KNN e LMinMAx, RDP e SobeDesce, gera o melhor desempenho de classificação?

Como medida de desempenho foi escolhida a AUC dos classificadores induzidos, uma vez que se trata de um problema de classes desbalance-

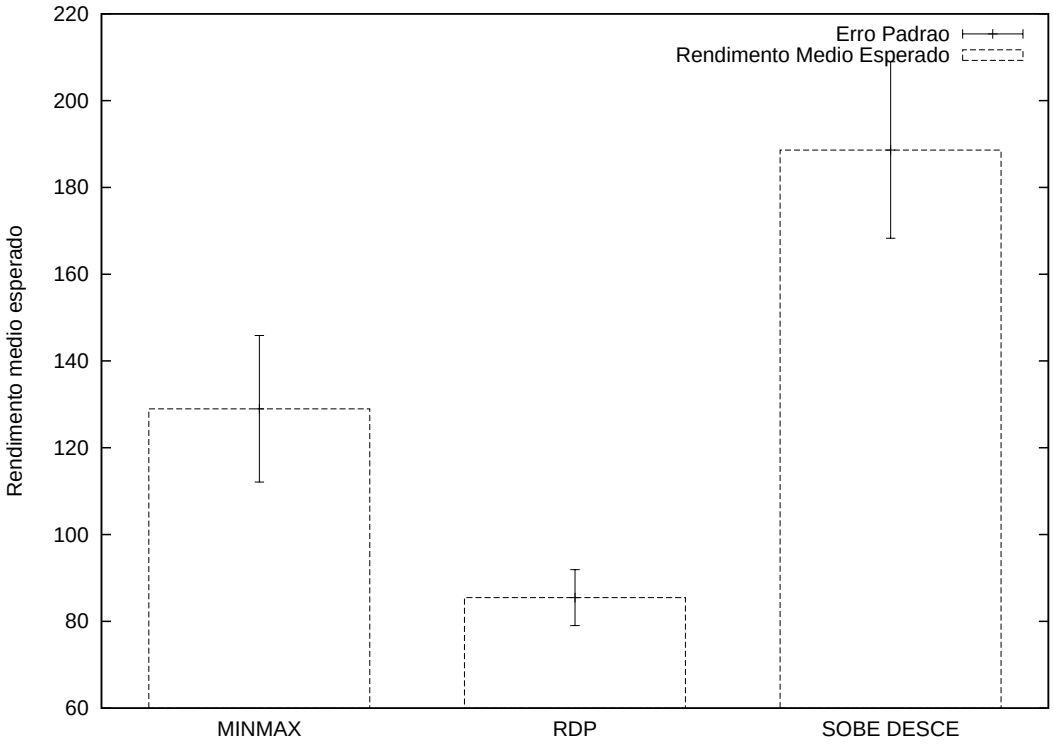

Figura 5.4: Gráfico do rendimento médio esperado (limite superior) e erro padrão por método de rotulação

adas. Além disso, a média da AUC foi ponderada pela quantidade exemplos por classe. Os valores de AUC dos classificadores induzidos com cada combinação de conjunto de dados, algoritmo e método de rotulação foram submetidos ao teste estatístico de *Friedman*, para verificar a existência de diferença significativa.

A estatística F resultante do teste de *Friedman* é 1002.49. Considerando que, com 5 e 345 graus de liberdade e 5% de significância, o valor crítico da estatística F é 2.24, a hipótese nula de que todas as combinações são iguais pode ser rejeitada, ou seja, pode-se dizer que há diferença significativa entre os resultados.

Para realizar o teste estatístico, os valores de AUC são ordenados pelo *rank* por conjunto de dados, conforme ilustrado na Tabela 5.13. Na primeira coluna tem-se o *split* do conjunto de dados com o nome da ação, seguido das colunas referentes à AUC do classificador de cada combinação de algoritmo e método de rotulação. Observe que, nas últimas colunas, o valor do *rank* aparece entre parenteses ao lado dos valores de AUC.

As médias de *rank* das combinações (quanto menor, melhor) são ilustradas na Figura 5.5. Observe que o LMINMAX é ordenado pelo *rank* em primeiro lugar, tanto para SVM quanto para KNN, seguido pelo RDP e SOBEDESCE, também com SVM e KNN nessa ordem.

Conjunto de dados	KNN			SVM		
	LMinMax	RDP	SobeDesce	LMinMax	RDP	SobeDesce
2008-01 2009-12 2010-12 ABEV3	0.642(2.0)	0.514(4.0)	0.248(5.0)	0.743(1.0)	0.517(3.0)	0.247(6.0)
2008-01 2009-12 2010-12 ABEV3	0.642(2.0)	0.514(4.0)	0.248(5.0)	0.743(1.0)	0.517(3.0)	0.247(6.0)
2008-01 2009-12 2010-12 BBAS3	0.733(2.0)	0.589(3.0)	0.261(5.0)	0.828(1.0)	0.560(4.0)	0.258(6.0)
2008-01 2009-12 2010-12 BBDC4	0.843(2.0)	0.445(4.0)	0.225(6.0)	0.844(1.0)	0.497(3.0)	0.260(5.0)
2008-01 2009-12 2010-12 BRFS3	0.788(2.0)	0.506(3.0)	0.251(5.0)	0.806(1.0)	0.472(4.0)	0.245(6.0)
2008-01 2009-12 2010-12 ITSA4	0.743(2.0)	0.528(4.0)	0.244(6.0)	0.829(1.0)	0.546(3.0)	0.271(5.0)
2008-01 2009-12 2010-12 ITUB4	0.691(2.0)	0.547(3.0)	0.254(5.0)	0.745(1.0)	0.506(4.0)	0.250(6.0)
2008-01 2009-12 2010-12 PETR3	0.681(2.0)	0.547(4.0)	0.245(6.0)	0.778(1.0)	0.561(3.0)	0.257(5.0)
2008-01 2009-12 2010-12 PETR4	0.735(2.0)	0.502(4.0)	0.241(6.0)	0.790(1.0)	0.599(3.0)	0.264(5.0)
2008-01 2009-12 2010-12 VALE3	0.704(2.0)	0.507(4.0)	0.257(6.0)	0.801(1.0)	0.566(3.0)	0.275(5.0)
2008-01 2009-12 2010-12 VALE5	0.690(2.0)	0.557(3.0)	0.258(5.0)	0.800(1.0)	0.516(4.0)	0.257(6.0)
2008-05 2010-04 2011-04 ABEV3	0.668(2.0)	0.510(3.0)	0.263(5.0)	0.693(1.0)	0.412(4.0)	0.253(6.0)
2008-05 2010-04 2011-04 BBAS3	0.797(2.0)	0.503(4.0)	0.251(5.0)	0.873(1.0)	0.559(3.0)	0.239(6.0)
2008-05 2010-04 2011-04 BBDC4	0.917(1.0)	0.528(3.0)	0.265(6.0)	0.894(2.0)	0.497(4.0)	0.266(5.0)
2008-05 2010-04 2011-04 BRFS3	0.679(2.0)	0.449(4.0)	0.263(5.0)	0.731(1.0)	0.465(3.0)	0.246(6.0)
2008-05 2010-04 2011-04 ITSA4	0.764(2.0)	0.503(4.0)	0.270(5.0)	0.830(1.0)	0.509(3.0)	0.241(6.0)
2008-05 2010-04 2011-04 ITUB4	0.789(2.0)	0.515(3.0)	0.265(5.0)	0.792(1.0)	0.500(4.0)	0.243(6.0)
2008-05 2010-04 2011-04 PETR3	0.715(2.0)	0.512(4.0)	0.211(6.0)	0.813(1.0)	0.567(3.0)	0.244(5.0)
2008-05 2010-04 2011-04 PETR4	0.737(2.0)	0.537(4.0)	0.245(6.0)	0.808(1.0)	0.582(3.0)	0.260(5.0)
2008-05 2010-04 2011-04 VALE3	0.743(2.0)	0.524(3.0)	0.260(5.0)	0.771(1.0)	0.516(4.0)	0.257(6.0)
2008-05 2010-04 2011-04 VALE5	0.706(2.0)	0.523(4.0)	0.239(6.0)	0.765(1.0)	0.539(3.0)	0.252(5.0)
2009-01 2010-12 2011-12 ABEV3	0.727(2.0)	0.536(3.0)	0.246(6.0)	0.773(1.0)	0.496(4.0)	0.256(5.0)
2009-01 2010-12 2011-12 BBAS3	0.783(2.0)	0.493(4.0)	0.241(6.0)	0.890(1.0)	0.509(3.0)	0.255(5.0)
2009-01 2010-12 2011-12 BBDC4	0.796(2.0)	0.571(3.0)	0.249(6.0)	0.865(1.0)	0.536(4.0)	0.253(5.0)
2009-01 2010-12 2011-12 BRFS3	0.769(2.0)	0.550(4.0)	0.256(5.0)	0.775(1.0)	0.576(3.0)	0.225(6.0)
2009-01 2010-12 2011-12 ITSA4	0.799(2.0)	0.560(3.0)	0.249(6.0)	0.840(1.0)	0.556(4.0)	0.255(5.0)
2009-01 2010-12 2011-12 ITUB4	0.746(2.0)	0.463(4.0)	0.251(6.0)	0.791(1.0)	0.539(3.0)	0.269(5.0)
2009-01 2010-12 2011-12 PETR3	0.782(2.0)	0.573(3.0)	0.295(5.0)	0.844(1.0)	0.510(4.0)	0.223(6.0)
2009-01 2010-12 2011-12 PETR4	0.693(2.0)	0.577(4.0)	0.247(6.0)	0.803(1.0)	0.593(3.0)	0.268(5.0)
2009-01 2010-12 2011-12 VALE3	0.719(2.0)	0.544(4.0)	0.262(6.0)	0.798(1.0)	0.549(3.0)	0.267(5.0)
2009-01 2010-12 2011-12 VALE5	0.752(2.0)	0.500(4.0)	0.275(5.0)	0.802(1.0)	0.509(3.0)	0.261(6.0)
2009-08 2011-07 2012-07 ABEV3	0.683(1.0)	0.499(3.0)	0.260(5.0)	0.671(2.0)	0.397(4.0)	0.238(6.0)
2009-08 2011-07 2012-07 BBAS3	0.714(2.0)	0.555(3.0)	0.236(6.0)	0.817(1.0)	0.455(4.0)	0.254(5.0)
2009-08 2011-07 2012-07 BBDC4	0.798(1.0)	0.512(3.0)	0.263(5.0)	0.784(2.0)	0.443(4.0)	0.214(6.0)
2009-08 2011-07 2012-07 BRFS3	0.733(2.0)	0.489(4.0)	0.248(6.0)	0.736(1.0)	0.544(3.0)	0.263(5.0)
2009-08 2011-07 2012-07 ITSA4	0.786(2.0)	0.617(3.0)	0.280(5.0)	0.828(1.0)	0.572(4.0)	0.268(6.0)
2009-08 2011-07 2012-07 ITUB4	0.706(2.0)	0.539(3.0)	0.265(6.0)	0.789(1.0)	0.508(4.0)	0.267(5.0)
2009-08 2011-07 2012-07 PETR3	0.824(2.0)	0.520(3.0)	0.246(5.0)	0.837(1.0)	0.492(4.0)	0.234(6.0)
2009-08 2011-07 2012-07 PETR4	0.743(2.0)	0.500(3.0)	0.243(5.0)	0.777(1.0)	0.399(4.0)	0.238(6.0)
2009-08 2011-07 2012-07 VALE3	0.725(2.0)	0.511(4.0)	0.242(6.0)	0.774(1.0)	0.549(3.0)	0.248(5.0)
2009-08 2011-07 2012-07 VALE5	0.760(2.0)	0.519(4.0)	0.257(6.0)	0.805(1.0)	0.608(3.0)	0.265(5.0)
2010-01 2011-12 2012-12 ABEV3	0.606(2.0)	0.501(3.0)	0.261(5.0)	0.746(1.0)	0.480(4.0)	0.231(6.0)
2010-01 2011-12 2012-12 BBAS3	0.728(2.0)	0.528(3.0)	0.246(6.0)	0.773(1.0)	0.524(4.0)	0.263(5.0)
2010-01 2011-12 2012-12 BBDC4	0.726(2.0)	0.559(3.0)	0.257(5.0)	0.735(1.0)	0.521(4.0)	0.252(6.0)
2010-01 2011-12 2012-12 BRFS3	0.713(2.0)	0.469(4.0)	0.245(6.0)	0.784(1.0)	0.478(3.0)	0.248(5.0)
2010-01 2011-12 2012-12 ITSA4	0.777(2.0)	0.512(4.0)	0.261(6.0)	0.853(1.0)	0.533(3.0)	0.278(5.0)
2010-01 2011-12 2012-12 ITUB4	0.817(2.0)	0.517(4.0)	0.242(6.0)	0.864(1.0)	0.528(3.0)	0.254(5.0)
2010-01 2011-12 2012-12 PETR3	0.806(2.0)	0.514(4.0)	0.253(5.0)	0.848(1.0)	0.574(3.0)	0.210(6.0)
2010-01 2011-12 2012-12 PETR4	0.713(2.0)	0.543(4.0)	0.248(6.0)	0.816(1.0)	0.570(3.0)	0.257(5.0)
2010-01 2011-12 2012-12 VALE3	0.708(2.0)	0.494(4.0)	0.261(6.0)	0.812(1.0)	0.509(3.0)	0.266(5.0)
2010-01 2011-12 2012-12 VALE5	0.763(2.0)	0.528(4.0)	0.289(6.0)	0.818(1.0)	0.585(3.0)	0.307(5.0)
2010-10 2012-09 2013-09 ABEV3	0.596(2.0)	0.501(3.0)	0.255(6.0)	0.649(1.0)	0.491(4.0)	0.256(5.0)
2010-10 2012-09 2013-09 BBAS3	0.750(2.0)	0.496(4.0)	0.270(5.0)	0.798(1.0)	0.549(3.0)	0.261(6.0)
2010-10 2012-09 2013-09 BBDC4	0.760(2.0)	0.480(3.0)	0.244(5.0)	0.832(1.0)	0.460(4.0)	0.221(6.0)
2010-10 2012-09 2013-09 BRFS3	0.559(2.0)	0.474(3.0)	0.260(6.0)	0.731(1.0)	0.460(4.0)	0.270(5.0)
2010-10 2012-09 2013-09 ITSA4	0.801(2.0)	0.590(3.0)	0.258(6.0)	0.862(1.0)	0.561(4.0)	0.270(5.0)
2010-10 2012-09 2013-09 ITUB4	0.838(2.0)	0.576(4.0)	0.261(6.0)	0.896(1.0)	0.600(3.0)	0.268(5.0)
2010-10 2012-09 2013-09 PETR3	0.728(2.0)	0.530(3.0)	0.244(6.0)	0.745(1.0)	0.458(4.0)	0.258(5.0)
2010-10 2012-09 2013-09 PETR4	0.823(1.0)	0.524(4.0)	0.264(5.0)	0.812(2.0)	0.538(3.0)	0.230(6.0)
2010-10 2012-09 2013-09 VALE3	0.716(2.0)	0.489(4.0)	0.255(5.0)	0.850(1.0)	0.508(3.0)	0.252(6.0)
2010-10 2012-09 2013-09 VALE5	0.644(2.0)	0.563(3.0)	0.236(6.0)	0.778(1.0)	0.540(4.0)	0.262(5.0)
2011-01 2012-12 2013-12 ABEV3	0.663(2.0)	0.518(4.0)	0.279(6.0)	0.758(1.0)	0.520(3.0)	0.291(5.0)
2011-01 2012-12 2013-12 BBAS3	0.773(2.0)	0.527(4.0)	0.242(5.0)	0.800(1.0)	0.554(3.0)	0.239(6.0)
2011-01 2012-12 2013-12 BBDC4	0.780(2.0)	0.566(3.0)	0.250(6.0)	0.840(1.0)	0.535(4.0)	0.275(5.0)
2011-01 2012-12 2013-12 BRFS3	0.613(3.0)	0.523(4.0)	0.251(6.0)	0.684(1.0)	0.630(2.0)	0.253(5.0)
2011-01 2012-12 2013-12 ITSA4	0.801(2.0)	0.527(3.0)	0.243(6.0)	0.885(1.0)	0.511(4.0)	0.284(5.0)
2011-01 2012-12 2013-12 ITUB4	0.704(2.0)	0.592(4.0)	0.244(6.0)	0.865(1.0)	0.620(3.0)	0.257(5.0)
2011-01 2012-12 2013-12 PETR3	0.696(2.0)	0.497(4.0)	0.253(6.0)	0.774(1.0)	0.507(3.0)	0.267(5.0)
2011-01 2012-12 2013-12 PETR4	0.846(2.0)	0.575(3.0)	0.288(5.0)	0.856(1.0)	0.554(4.0)	0.226(6.0)
2011-01 2012-12 2013-12 VALE3	0.754(2.0)	0.527(3.0)	0.235(6.0)	0.823(1.0)	0.497(4.0)	0.251(5.0)
2011-01 2012-12 2013-12 VALE5	0.684(2.0)	0.549(4.0)	0.233(6.0)	0.811(1.0)	0.566(3.0)	0.240(5.0)
Rank médio	1.957	3.543	5.6	1.057	3.443	5.4

Tabela 5.13: Ordenação da AUC por *rank*, por conjunto de dados, para cálculo da diferença significativa entre as combinações de algoritmos e métodos de rotulação

Após o teste de Friedman detectar a existência de diferença significativa entre as combinações, pode-se prosseguir com a análise *post-hoc* de *Nemenyi* para verificar entre quais combinações realmente existe diferença. O teste de *Nemenyi* mostra que as médias de *rank* com diferenças maiores que 0.9 podem ser consideradas significativamente diferentes. Na Figura 5.6 é ilustrado um gráfico de diferença crítica, no qual o eixo prin-

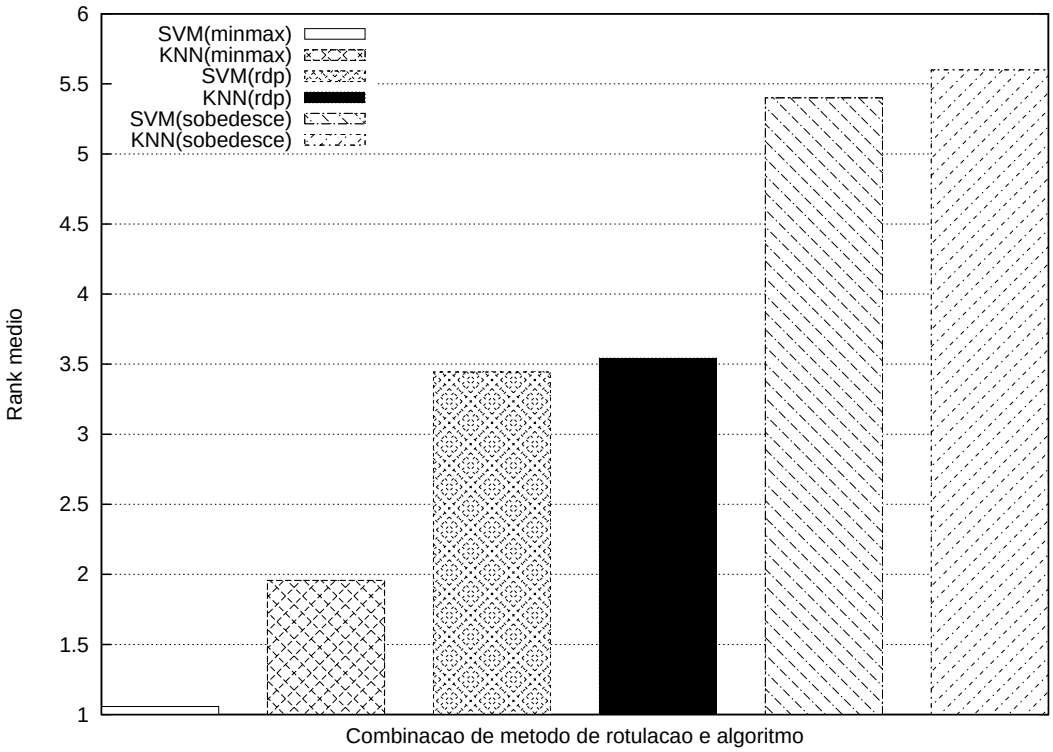

Figura 5.5: *Rank* médio do teste estatístico de Friedman entre os valores de AUC das combinações de algoritmos e métodos de rotulação

cipal representa as médias de *rank*. Observe os traços horizontais de tamanho 0.9, que "ligam" as combinações cuja diferença não é significativa.

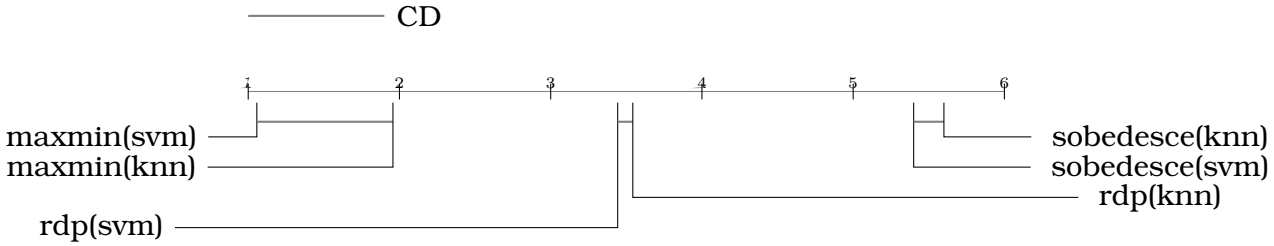

Figura 5.6: Gráfico de diferença crítica entre os valores de AUC das combinações de algoritmos e métodos de rotulação

Com base no resultado da análise *post-hoc* de *Nemenyi*, as combinações podem ser avaliadas em pares. Na Tabela 5.14 são ilustradas todos os pares de combinações e se entre elas há empate ou se uma é melhor ou pior que a outra. Assim, pode-se concluir que o método de rotulação com melhor desempenho, com diferença significativa, entre LMINMAX, RDP e SOBEDESCE, é o LMINMAX. Além disso, pode-se concluir que o desempenho de classificação do SVM e KNN é similar, porque para os três métodos de rotulação houve empate entre os dois algoritmos.

3. Melhor desempenho na classificação indica maior lucro na simulação?

Método		LMinMAx		RDP		SobeDesce	
	Algoritmo	SVM	KNN	SVM	KNN	SVM	KNN
LMinMAx	SVM	empate	empate	melhor	melhor	melhor	melhor
	KNN	empate	empate	melhor	melhor	melhor	melhor
RDP	SVM	pior	pior	empate	empate	melhor	melhor
	KNN	pior	pior	empate	empate	melhor	melhor
SobeDesce	SVM	pior	pior	pior	pior	empate	empate
	KNN	pior	pior	pior	pior	empate	empate

Tabela 5.14: Resultado (empate, melhor ou pior) da diferença significativa entre os valores de AUC dos classificadores induzidos com as combinações de algoritmos e métodos de rotulação

O teste estatístico de Friedman também foi aplicado sobre os rendimentos resultantes da simulação, das diferentes combinações de algoritmo e método de rotulação, para verificar se há diferença significativa. A estatística F resultante foi de 34.70, com 5 e 345 graus de liberdade, para 5% de significância. Na Tabela 5.15 são ilustrados os valores de *rank* obtidos do teste e, na Figura 5.7 os *rank* médios ordenados. Observe que o LMinMAx é ordenado em primeiro lugar, tanto para SVM quanto para KNN, seguido pelo RDP e SobeDesce, também com SVM e KNN nessa ordem.

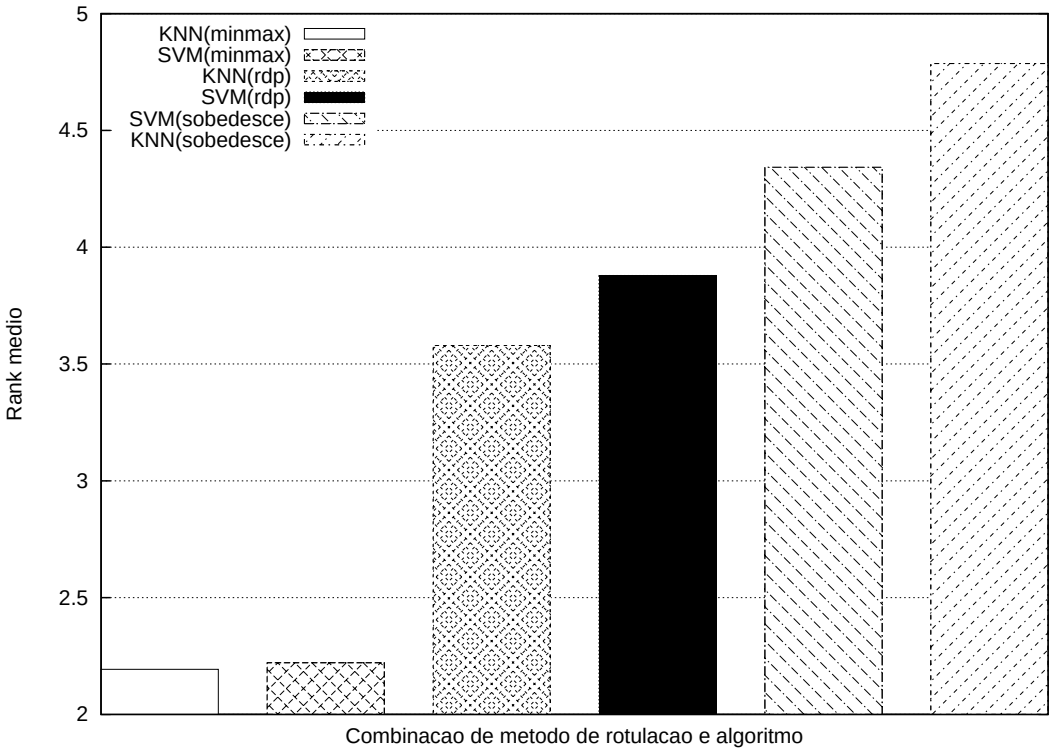

Figura 5.7: *Rank* médio do teste estatístico entre os valores de rendimento na simulação de todas as combinações de algoritmos e métodos de rotulação da avaliação experimental

Com a análise *post-hoc* de *Nemenyi* pode-se verificar resultado semelhante ao alcançado com os valores de AUC da questão anterior. De

Conjunto de dados	KNN			SVM		
	LMinMax	RDP	SobeDesce	LMinMax	RDP	SobeDesce
2008-01 2009-12 2010-12 ABEV3	85.251(1.0)	13.844(3.5)	0.889(6.0)	77.760(2.0)	13.844(3.5)	9.633(5.0)
2008-01 2009-12 2010-12 BBAS3	15.277(2.0)	-20.642(5.0)	-29.706(6.0)	16.851(1.0)	-10.619(4.0)	4.032(3.0)
2008-01 2009-12 2010-12 BBDC4	47.315(2.0)	-12.878(3.0)	-27.849(4.0)	52.014(1.0)	-30.563(5.0)	-43.523(6.0)
2008-01 2009-12 2010-12 BRFS3	7.573(3.0)	5.412(4.0)	-12.808(6.0)	9.532(2.0)	11.700(1.0)	-12.808(5.0)
2008-01 2009-12 2010-12 ITSA4	11.170(3.0)	-15.842(5.0)	-16.299(6.0)	27.003(1.0)	5.472(4.0)	11.262(2.0)
2008-01 2009-12 2010-12 ITUB4	0.000(2.0)	-7.774(3.0)	-35.042(6.0)	18.286(1.0)	-10.986(4.0)	-12.577(5.0)
2008-01 2009-12 2010-12 PETR3	-10.237(3.0)	-34.117(6.0)	-28.302(5.0)	-1.580(1.0)	-21.842(4.0)	-7.896(2.0)
2008-01 2009-12 2010-12 PETR4	-17.988(3.0)	-11.191(2.0)	-21.778(4.0)	0.881(1.0)	-34.470(6.0)	-28.810(5.0)
2008-01 2009-12 2010-12 VALE3	3.598(2.0)	8.445(1.0)	-19.864(6.0)	-10.153(5.0)	3.463(3.0)	0.000(4.0)
2008-01 2009-12 2010-12 VALE5	3.021(3.0)	-0.851(4.0)	-12.444(5.0)	9.179(2.0)	22.502(1.0)	-16.670(6.0)
2008-05 2010-04 2011-04 ABEV3	4.354(3.0)	41.745(1.0)	-6.728(6.0)	0.000(5.0)	5.230(2.0)	3.255(4.0)
2008-05 2010-04 2011-04 BBAS3	28.624(1.0)	3.961(4.0)	-13.375(5.0)	9.159(2.0)	8.497(3.0)	-39.800(6.0)
2008-05 2010-04 2011-04 BBDC4	51.840(1.0)	6.157(4.0)	-11.079(5.0)	43.044(2.0)	22.801(3.0)	-45.348(6.0)
2008-05 2010-04 2011-04 BRFS3	11.888(2.0)	0.830(3.0)	-22.053(6.0)	16.403(1.0)	-0.723(4.0)	-22.053(5.0)
2008-05 2010-04 2011-04 ITSA4	32.076(1.0)	-10.949(6.0)	0.950(4.0)	24.373(2.0)	7.026(3.0)	-7.500(5.0)
2008-05 2010-04 2011-04 ITUB4	34.422(1.0)	-9.654(5.0)	-7.329(4.0)	18.652(2.0)	-20.306(6.0)	0.000(3.0)
2008-05 2010-04 2011-04 PETR3	-14.055(3.0)	-22.879(4.0)	-40.169(5.0)	7.824(1.0)	-5.737(2.0)	-51.089(6.0)
2008-05 2010-04 2011-04 PETR4	-1.387(3.0)	29.530(1.0)	-19.315(5.0)	-5.199(4.0)	5.550(2.0)	-39.743(6.0)
2008-05 2010-04 2011-04 VALE3	6.634(2.0)	-17.859(5.0)	-8.525(4.0)	2.607(3.0)	10.826(1.0)	-20.711(6.0)
2008-05 2010-04 2011-04 VALE5	14.962(1.0)	-5.618(4.0)	-3.220(3.0)	6.045(2.0)	-7.986(5.0)	-18.398(6.0)
2009-01 2010-12 2011-12 ABEV3	68.108(1.0)	6.007(3.5)	-8.395(6.0)	15.477(2.0)	6.007(3.5)	-1.112(5.0)
2009-01 2010-12 2011-12 BBAS3	16.786(1.0)	-31.873(6.0)	-21.191(4.0)	-10.564(3.0)	-22.748(5.0)	-8.796(2.0)
2009-01 2010-12 2011-12 BBDC4	8.730(1.0)	-6.620(5.0)	-42.810(6.0)	7.892(2.0)	-0.095(4.0)	0.000(3.0)
2009-01 2010-12 2011-12 BRFS3	-4.779(5.0)	-0.928(4.0)	1.532(3.0)	7.646(1.0)	-15.629(6.0)	3.490(2.0)
2009-01 2010-12 2011-12 ITSA4	24.983(1.0)	5.016(2.0)	-23.061(5.0)	4.626(3.0)	-42.153(6.0)	-9.482(4.0)
2009-01 2010-12 2011-12 ITUB4	19.869(1.0)	-21.443(6.0)	-17.731(4.0)	1.603(3.0)	10.291(2.0)	-18.737(5.0)
2009-01 2010-12 2011-12 PETR3	-10.261(4.0)	8.408(1.0)	-6.238(3.0)	-19.809(5.0)	-34.198(6.0)	-1.990(2.0)
2009-01 2010-12 2011-12 PETR4	-17.572(2.0)	-19.965(4.0)	-25.000(6.0)	-19.832(3.0)	11.598(1.0)	-25.000(5.0)
2009-01 2010-12 2011-12 VALE3	-18.339(3.0)	-14.527(2.0)	-35.011(5.0)	-5.394(1.0)	-45.191(6.0)	-24.178(4.0)
2009-01 2010-12 2011-12 VALE5	-17.931(4.0)	-7.191(3.0)	-2.723(1.0)	-7.152(2.0)	-30.188(6.0)	-18.900(5.0)
2009-08 2011-07 2012-07 ABEV3	0.423(3.0)	-3.649(5.0)	20.409(1.0)	0.000(4.0)	-3.816(6.0)	8.385(2.0)
2009-08 2011-07 2012-07 BBAS3	-8.849(1.0)	-9.202(2.0)	-42.072(5.0)	-10.560(3.0)	-12.080(4.0)	-46.104(6.0)
2009-08 2011-07 2012-07 BBDC4	38.437(1.0)	3.465(3.0)	-22.417(6.0)	21.602(2.0)	0.682(4.0)	-8.405(5.0)
2009-08 2011-07 2012-07 BRFS3	17.517(2.0)	-12.069(5.0)	-32.667(6.0)	12.902(3.0)	-4.327(4.0)	20.835(1.0)
2009-08 2011-07 2012-07 ITSA4	20.342(1.0)	-0.553(4.0)	-27.422(5.0)	16.745(2.0)	10.291(3.0)	-35.197(6.0)
2009-08 2011-07 2012-07 ITUB4	27.916(3.0)	-3.297(5.0)	-15.516(6.0)	31.334(2.0)	37.058(1.0)	-0.214(4.0)
2009-08 2011-07 2012-07 PETR3	14.983(2.0)	13.001(3.0)	-15.598(4.0)	18.312(1.0)	-16.433(5.0)	-39.169(6.0)
2009-08 2011-07 2012-07 PETR4	3.465(2.0)	-13.659(4.0)	-31.068(6.0)	6.798(1.0)	-13.659(3.0)	-24.400(5.0)
2009-08 2011-07 2012-07 VALE3	-11.518(3.0)	-22.440(4.0)	-48.008(6.0)	-6.824(2.0)	-35.008(5.0)	0.000(1.0)
2009-08 2011-07 2012-07 VALE5	16.359(1.0)	-17.839(3.0)	-59.630(6.0)	4.872(2.0)	-30.531(5.0)	-24.970(4.0)
2010-01 2011-12 2012-12 ABEV3	0.000(4.0)	16.112(2.5)	-8.672(6.0)	60.580(1.0)	16.112(2.5)	-1.394(5.0)
2010-01 2011-12 2012-12 BBAS3	-10.578(3.0)	14.096(1.0)	-18.847(5.0)	-4.766(2.0)	-16.585(4.0)	-26.100(6.0)
2010-01 2011-12 2012-12 BBDC4	7.671(3.0)	-15.810(6.0)	11.644(2.0)	12.901(1.0)	4.766(4.0)	-1.970(5.0)
2010-01 2011-12 2012-12 BRFS3	2.976(2.0)	5.698(1.0)	-22.850(6.0)	2.646(3.0)	-18.115(5.0)	-0.128(4.0)
2010-01 2011-12 2012-12 ITSA4	4.631(1.0)	1.558(2.0)	-21.574(6.0)	-7.325(3.0)	-13.348(4.0)	-20.488(5.0)
2010-01 2011-12 2012-12 ITUB4	3.240(1.0)	0.619(3.0)	-23.512(6.0)	1.678(2.0)	-17.859(4.0)	-20.977(5.0)
2010-01 2011-12 2012-12 PETR3	-3.787(2.0)	-12.123(3.0)	-38.812(6.0)	-3.322(1.0)	-28.537(5.0)	-19.712(4.0)
2010-01 2011-12 2012-12 PETR4	-20.802(5.0)	-14.158(3.0)	4.444(1.0)	-16.175(4.0)	-7.736(2.0)	-36.200(6.0)
2010-01 2011-12 2012-12 VALE3	18.968(1.0)	-28.466(5.0)	3.206(3.0)	4.899(2.0)	-9.977(4.0)	-29.940(6.0)
2010-01 2011-12 2012-12 VALE5	13.050(1.0)	-27.301(6.0)	-13.787(4.0)	11.595(2.0)	-13.842(5.0)	5.226(3.0)
2010-10 2012-09 2013-09 ABEV3	0.000(4.5)	13.389(2.0)	-27.650(6.0)	0.000(4.5)	18.280(1.0)	5.600(3.0)
2010-10 2012-09 2013-09 BBAS3	14.388(1.0)	4.413(3.0)	-5.062(5.0)	5.734(2.0)	-5.554(6.0)	0.412(4.0)
2010-10 2012-09 2013-09 BBDC4	22.167(1.0)	3.693(4.0)	10.197(3.0)	21.740(2.0)	-21.448(6.0)	-14.612(5.0)
2010-10 2012-09 2013-09 BRFS3	0.000(3.0)	8.182(2.0)	-8.500(6.0)	-0.140(4.0)	11.396(1.0)	-1.799(5.0)
2010-10 2012-09 2013-09 ITSA4	-0.670(2.0)	-14.572(4.0)	-10.587(3.0)	33.836(1.0)	-15.361(5.0)	-16.862(6.0)
2010-10 2012-09 2013-09 ITUB4	28.764(2.0)	-2.164(3.0)	-21.816(6.0)	52.894(1.0)	-12.636(4.0)	-19.205(5.0)
2010-10 2012-09 2013-09 PETR3	-31.140(4.0)	-4.822(2.0)	-33.142(6.0)	-21.615(3.0)	-1.988(1.0)	-33.142(5.0)
2010-10 2012-09 2013-09 PETR4	-0.544(2.0)	2.350(1.0)	-24.061(5.0)	-9.063(4.0)	-5.765(3.0)	-39.606(6.0)
2010-10 2012-09 2013-09 VALE3	-4.708(2.0)	-24.753(5.0)	-28.292(6.0)	-1.065(1.0)	-9.171(3.0)	-21.370(4.0)
2010-10 2012-09 2013-09 VALE5	-13.497(3.0)	-25.814(5.0)	-20.055(4.0)	-4.517(2.0)	-27.842(6.0)	0.000(1.0)
2011-01 2012-12 2013-12 ABEV3	17.856(1.0)	-3.790(6.0)	-3.615(4.0)	12.797(2.0)	-3.790(5.0)	-0.903(3.0)
2011-01 2012-12 2013-12 BBAS3	-11.029(4.0)	6.309(2.0)	-33.033(6.0)	-11.275(5.0)	17.907(1.0)	-7.820(3.0)
2011-01 2012-12 2013-12 BBDC4	5.635(3.0)	-9.559(5.0)	-0.087(4.0)	13.928(1.0)	-16.032(6.0)	8.439(2.0)
2011-01 2012-12 2013-12 BRFS3	12.420(1.0)	-7.052(4.0)	-28.297(6.0)	5.195(2.0)	-7.052(3.0)	-24.976(5.0)
2011-01 2012-12 2013-12 ITSA4	4.926(3.0)	15.989(1.0)	-27.790(6.0)	9.193(2.0)	0.968(4.0)	-13.906(5.0)
2011-01 2012-12 2013-12 ITUB4	5.638(2.0)	2.309(3.0)	-19.208(5.0)	32.308(1.0)	-6.566(4.0)	-24.893(6.0)
2011-01 2012-12 2013-12 PETR3	-21.503(4.0)	-5.680(1.0)	-32.998(5.0)	-13.488(3.0)	-41.442(6.0)	-13.416(2.0)
2011-01 2012-12 2013-12 PETR4	10.659(1.0)	-24.626(5.0)	-8.846(2.0)	-9.894(3.0)	-27.148(6.0)	-9.947(4.0)
2011-01 2012-12 2013-12 VALE3	8.475(2.0)	-17.626(6.0)	-12.550(4.0)	10.963(1.0)	-15.300(5.0)	3.024(3.0)
2011-01 2012-12 2013-12 VALE5	-5.686(1.0)	-31.412(6.0)	-8.857(3.0)	-8.337(2.0)	-11.714(4.0)	-29.034(5.0)
Rank médio	2.193	3.579	4.786	2.221	3.879	4.343

Tabela 5.15: Ordenação do rendimento na simulação por *rank*, por conjunto de dados, para cálculo da diferença significativa entre as combinações de algoritmos e métodos de rotulação

acordo com a estatística Nemenyi, diferenças maiores que 0.9 entre os *rank* médios podem indicam diferença significativa. Na Figura 5.8 é ilustrado um gráfico de diferença crítica para os resultados de simulação, no qual o eixo principal representa as médias de *rank*. Observe os traços horizontais de tamanho 0.9, que "ligam" as combinações cuja diferença não é significativa.

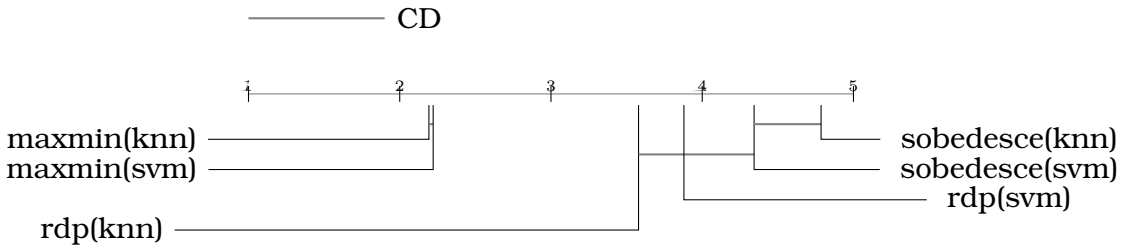

Figura 5.8: Gráfico de diferença crítica entre os valores de rendimento da simulação das combinações de algoritmos e métodos de rotulação

Método		LMinMAx		RDP		SobeDesce	
	Algoritmo	SVM	KNN	SVM	KNN	SVM	KNN
LMinMAx	KNN	empate	empate	melhor	melhor	melhor	melhor
	SVM	empate	empate	melhor	melhor	melhor	melhor
RDP	KNN	pior	pior	empate	empate	empate	melhor
	SVM	pior	pior	empate	empate	empate	melhor
SobeDesce	KNN	pior	pior	empate	empate	empate	empate
	SVM	pior	pior	pior	pior	empate	empate

Tabela 5.16: Resultado (empate, melhor ou pior) da diferença significativa entre os rendimentos na simulação das combinações de algoritmos e métodos de rotulação

Como o teste estatístico mostrou que os melhores resultados de classificação geram maior rendimento na simulação, e que o método de rotulação LMinMAx supera os demais, o restante deste capítulo será focado nos resultados para o LMinMAx. Por outro lado, como houve empate entre os algoritmos SVM e KNN, ambos serão considerados na apresentação dos resultados.

Além do teste estatístico, a correlação entre a acurácia de classificação e o lucro também foi calculada. Considerando o desbalanceamento das classes, a correlação foi medida separadamente por classe e apenas para as classes relevantes. Os resultados mostraram correlação entre o rendimento e a classe MIN de 0.26, e para a classe MAX, 0.06.

Então, pode-se concluir que o rendimento na simulação está relacionado ao resultado da classificação (AUC) e, mais especificamente, correlacionado com a acurácia de classificação, nesse caso, da classe MIN.

4. A rentabilidade da abordagem proposta é superior às estratégias de investimento atualmente realizadas no mercado financeiro?

   Para comparar a abordagem proposta com o desempenho humano, foram consideradas as rentabilidades de duas estratégias financeiras conhecidas: aplicação em poupança e carteiras recomendadas.

   Como a avaliação das carteiras é realizada anualmente, foram considerados os conjuntos de dados cujos *split* compreendem o conjunto de teste

nos anos de 2011, 2012 e 2013. Além disso, considerando que o desempenho das corretoras varia anualmente, a abordagem proposta é comparada não só com a melhor de cada ano, mas também com a média das dez melhores.

A Tabela 5.18 mostra as dez melhores carteiras recomendadas de 2011, 2012 e 2013, retiradas do site http://exame.abril.com.br. Observe que, no ano de 2012, todas as carteiras mostradas tiveram rendimento favorável, em média 42.19%. Já nos anos de 2011 e 2013, o resultado foi bem diferente. Em 2011 apenas uma carteira obteve lucro, a Souza Barros, com 6.74%. As demais apresentam um prejuízo de, em média, -9.23%. Em 2013 o cenário foi mais equilibrado, as carteiras tiveram mais rendimento favorável. Entretanto o lucro foi baixo, de apenas 3.56%.

Quanto à abordagem proposta, o desempenho nos três anos se mostrou mais equilibrado. Na Tabela 5.17 é ilustrada a média de rendimento das dez ações do conjunto experimental. Observe que, em 2011, o desempenho foi de -2.55% com o SVM e 6.95% com o KNN. Nesse ano o desempenho do KNN supera as carteiras recomendadas, que alcançaram no máximo 6.74% e em média -9.23%. Em 2012, embora o rendimento médio tenha sido positivo, de 6.27% para o SVM e 1.53%, ficou bem abaixo do obtido com as carteiras no ano de 2012. Em 2013, a abordagem proposta com SVM mostra desempenho superior, com 4.13% contra os 3.55% em média das carteiras. Em 2013 o desempenho do KNN também foi positivo, com 2.73%.

Ação	2011		2012		2013	
	SVM	KNN	SVM	KNN	SVM	KNN
ABEV3	15.477	68.1084	60.5802	0	12.7967	17.856
BBAS3	-10.5643	16.7862	-4.7658	-10.5776	-11.2748	-11.029
BBDC4	7.8918	8.7302	12.9006	7.6706	13.9282	5.6349
BRFS3	7.6459	-4.7788	2.6463	2.9762	5.1951	12.4202
ITSA4	4.6264	24.9827	-7.3247	4.6311	9.1934	4.9264
ITUB4	1.6026	19.869	1.6784	3.2403	32.3084	5.6376
PETR3	-19.809	-10.2608	-3.3222	-3.7871	-13.4875	-21.5034
PETR4	-19.8316	-17.5723	-16.1746	-20.8023	-9.8944	10.6585
VALE3	-5.3945	-18.3386	4.8992	18.968	10.9628	8.4754
VALE5	-7.1515	-17.9308	11.5946	13.0495	-8.3374	-5.6861
Média	-2.55072	6.95952	6.2712	1.53687	4.13905	2.73905

Tabela 5.17: Rendimento médio por ação (em todos os conjuntos de dados) da abordagem proposta LMinMAx no conjunto experimental de dez ações

Na Figura 5.9 é ilustrada a comparação da média do conjunto experimental da abordagem proposta com as carteiras recomendadas nos três anos. Observe que, em 2011 e em 2013, a abordagem proposta tem rendimento superior, com KNN e SVM, respectivamente.

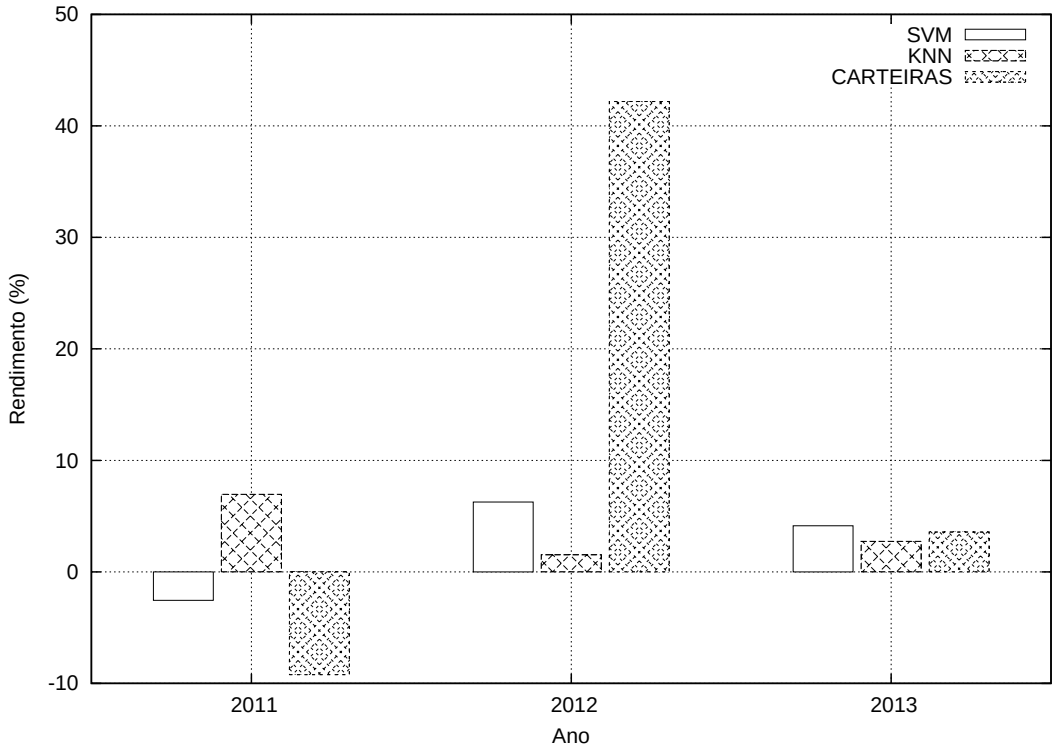

Figura 5.9: Rendimento da abordagem proposta LMINMAX comparada com as carteiras recomendadas nos anos de 2011, 2012 e 2013

Corretora	2011 (%)	Corretora	2012 (%)	Corretora	2013 (%)
Souza Barros	6.74	Geral	55.48	Ágora/Bradesco	15.30
HSBC	-3.41	Alpes/Wintrade	52.81	Geração Futuro	10.40
Planner	-5.38	Souza Barros	52.04	BTG Pactual	7.30
BTG Pactual	-6.60	Coinvalores	44.60	BB	2.43
WinTrade	-11.25	Fator	42.60	Pax	2.34
XP	-12.10	Rico/Octo	36.11	Rico/Octo	1.66
BB-BBI	-12.90	BI&P	35.26	Geral	0.93
Ativa	-13.60	XP	34.50	Concórdia	-0.35
Gradual	-16.30	BTG Pactual	34.40	HSBC	-2.20
Socopa	-17.53	Pax Corretora	34.08	Coinvalores	-2.22
Média 2011	-9.23	Média 2012	42.19	Média 2013	3.56

Tabela 5.18: Rendimento anual das TOP 10 carteiras recomendadas dos anos de 2011, 2012 e 2013

CAPÍTULO

6

Considerações Finais

As representações de classe normalmente utilizadas para problemas de predição de ações se baseiam na tendência de preço futura da ação e tendem a ser imediatistas. Consideram valores absolutos e classificam os movimentos de preço independente da intensidade de variação. Por outro lado, a rotulação com LMINMAX proposta por este trabalho, trata o atributo classe em um contexto maior, situando o preço da ação em relação ao período em que se encontra. Os conceitos de ponto máximo e mínimo permitem a identificação de dias relevantes na série temporal financeira, de modo que operações combinadas de compra e venda, nesses dias, alcancem desempenho bastante relevante.

A pesquisa realizada durante o desenvolvimento deste trabalho, bem como os resultados alcançados, permitem concluir que é possível representar séries temporais financeiras em tabelas atributo valor significativas para algoritmos de aprendizado de máquina supervisionado. Também acredita-se na possibilidade de classificar os preços quanto à variação em relação ao período em que se encontram, utilizando como informação apenas os preços históricos da ação até o dia que se deseja prever.

Como direções futuras deste trabalho, pode-se considerar algumas ideias surgidas no decorrer do desenvolvimento da pesquisa, mas que não foram abordadas por serem externas ao escopo do trabalho. Dentre as propostas sugere-se:

- melhorar a acurácia de classificação das classes relevantes MAX e MIN, utilizando a calibração dos classificadores
- promover a seleção de ações, a fim de investir naquelas cujo rendimento

tende a ser mais favorável

- tornar a abordagem proposta usável para recomendação automática de compras e venda de ações.
- aplicar o algoritmo proposto em conjuntos de ações restritos a setores de economia
- aplicar o algoritmo proposto em outros produtos financeiros, tais como mercado futuro e de opções, moeda e títulos
- analisar a correlação entre a acurácia de classificação (auc) e o rendimento nas operações de compra e venda

Referências Bibliográficas

(2013). Site uol economia.

Aha, D., Kibler, D., e Albert, M. (1991). Instance-based learning algorithms. Machine Learning, 6(1):37–66.

Atsalakis, G. S. e Valavanis, K. P. (2009). Surveying stock market forecasting techniques part ii: Soft computing methods. Expert Systems with Applications, 36(3, Part 2):5932 – 5941.

Bergstra, J. e Bengio, Y. (2012). Random search for hyper-parameter optimization. J. Mach. Learn. Res., 13:281–305.

Boyacioglu, M. A. e Avci, D. (2010). An adaptive network-based fuzzy inference system (anfis) for the prediction of stock market return: The case of the istanbul stock exchange. Expert Systems with Applications, 37(12):7908 – 7912.

Cao, L.-J. e Tay, F. E. H. (2003). Support vector machine with adaptive parameters in financial time series forecasting. Neural Networks, IEEE Transactions on, 14(6):1506–1518.

Chang, C.-C. e Lin, C.-J. (2011). Libsvm: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology (TIST), 2(3):27.

Chang, P.-C. e Fan, C.-Y. (2008). A hybrid system integrating a wavelet and tsk fuzzy rules for stock price forecasting. Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on, 38(6):802–815.

Chang, P.-C., Fan, C.-Y., e Liu, C.-H. (2009). Integrating a piecewise linear representation method and a neural network model for stock trading points prediction. Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on, 39(1):80–92.

Chen, Y., Mabu, S., Shimada, K., e Hirasawa, K. (2009). A genetic network programming with learning approach for enhanced stock trading model. Expert Systems with Applications, 36(10):12537–12546.

Choudhry, R. e Garg, K. (2008). A hybrid machine learning system for stock market forecasting. In Proceedings of World Academy of Science, Engineering and Technology, páginas 315–318.

Edson Takashi Matsubara, M. C. M. (2008). Teste. In Relações entre ranking, análise ROC e calibraccão em aprendizado de máquina.

Fawcett, T. (2006). An introduction to roc analysis. Pattern Recogn. Lett., 27(8):861–874.

Friedman, M. (2008). The use of ranks to avoid the assumption of nomality implicit in the analysis of variance. American Statistical Association, 32(200):675–701.

Fu, T.-c., Chung, F.-l., Luk, R., e Ng, C.-m. (2007). Stock time series pattern matching: Template-based vs. rule-based approaches. Eng. Appl. Artif. Intell., 20(3):347–364.

Hadavandi, E., Shavandi, H., e Ghanbari, A. (2010). Integration of genetic fuzzy systems and artificial neural networks for stock price forecasting. Knowledge-Based Systems, 23(8):800–808.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., e Witten, I. H. (2009). The weka data mining software: An update. SIGKDD Explor. Newsl., 11(1):10–18.

Hassan, M. R., Nath, B., e Kirley, M. (2007). A fusion model of hmm, ann and ga for stock market forecasting. Expert Systems with Applications, 33(1):171–180.

Hsu, C.-W., Chang, C.-C., Lin, C.-J., et al. (2003). A practical guide to support vector classification.

Huang, C.-L. e Tsai, C.-Y. (2009). A hybrid sofm-svr with a filter-based feature selection for stock market forecasting. Expert Systems with Applications, 36(2):1529–1539.

Huang, W., Nakamori, Y., e Wang, S.-Y. (2005). Forecasting stock market movement direction with support vector machine. Computers & Operations Research, 32(10):2513–2522.

jae Kim, K. (2003). Financial time series forecasting using support vector machines. Neurocomputing, 55(1 2):307 – 319. Support Vector Machines.

Kaastra, I. e Boyd, M. (1996). Designing a neural network for forecasting financial and economic time series. Neurocomputing, 10(3):215–236.

Kara, Y., Acar Boyacioglu, M., e Baykan, Ö. K. (2011). Predicting direction of stock price index movement using artificial neural networks and support vector machines: The sample of the istanbul stock exchange. Expert systems with Applications, 38(5):5311–5319.

Keogh, E. (2002). Exact indexing of dynamic time warping. In Proceedings of the 28th international conference on Very Large Data Bases, VLDB '02, páginas 406–417. VLDB Endowment.

Klassen, M. (2005). Investigation of some technical indexes in stock forecasting using neural networks. In WEC (5), páginas 75–79. Citeseer.

Lai, R. K., Fan, C.-Y., Huang, W.-H., e Chang, P.-C. (2009). Evolving and clustering fuzzy decision tree for financial time series data forecasting. Expert Systems with Applications, 36(2):3761–3773.

Lin, J., Keogh, E., Lonardi, S., e Patel, P. (2002). Finding motifs in time series. páginas 53–68.

ManChon, U. e Rasheed, K. (2010). A relative tendency based stock market prediction system. In Machine Learning and Applications (ICMLA), 2010 Ninth International Conference on, páginas 949–953. IEEE.

Martinez, L. C., da Hora, D. N., de M Palotti, J. R., Meira, W., e Pappa, G. L. (2009). From an artificial neural network to a stock market day-trading system: A case study on the bm&f bovespa. In Neural Networks, 2009. IJCNN 2009. International Joint Conference on, páginas 2006–2013. IEEE.

McCulloch, W. S. e Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. The Bulletin of Mathematical Biophysics, 5(4):115–133.

Mueen, A., Keogh, E., Zhu, Q., e Cash, S. (2009). Exact discovery of time series motifs. In SDM.

Quinlan, J. R. (1993). C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Rumelhart, D. E., Hinton, G. E., e Williams, R. J. (1988). Neurocomputing: Foundations of research. chapter Learning Representations by Back-propagating Errors, páginas 696–699. MIT Press, Cambridge, MA, USA.

Saffi, P. A. (2003). Análise técnica: sorte ou realidade? Revista Brasileira de Economia, 57(4):953–974.

Schölkopf, B. (2000). Statistical learning and kernel methods.

Smith, L. I. (2002). A tutorial on principal component analysis. http://www.cs.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf.

Tsai, C.-F. e Hsiao, Y.-C. (2010). Combining multiple feature selection methods for stock prediction: Union, intersection, and multi-intersection approaches. Decision Support Systems, 50(1):258–269.

Vapnik, V. N. (1995). The Nature of Statistical Learning Theory. Springer New York Inc., New York, NY, USA.

Wang, X., Mueen, A., Ding, H., Trajcevski, G., Scheuermann, P., e Keogh, E. (2013). Experimental comparison of representation methods and distance measures for time series data. Data Mining and Knowledge Discovery, 26(2):275–309.

Xi, X., Keogh, E., Shelton, C., Wei, L., e Ratanamahatana, C. A. (2006). Fast time series classification using numerosity reduction. In Proceedings of the 23rd international conference on Machine learning, páginas 1033–1040. ACM.

Xie, W., Yu, L., Xu, S., e Wang, S. (2006). A new method for crude oil price forecasting based on support vector machines. In Computational Science–ICCS 2006, páginas 444–451. Springer.

Yeh, C.-Y., Huang, C.-W., e Lee, S.-J. (2011). A multiple-kernel support vector regression approach for stock market price forecasting. Expert Systems with Applications, 38(3):2177–2186.