



Universidade Nova de Lisboa

Faculdade de Ciências e Tecnologia

Licenciatura em Engenharia Informática

Sistema de Apoio à Decisão para Investimento na Bolsa de Valores usando Data Mining

Projecto de Fim de Curso na Novabase Business Intelligence

Autores:

José Carlos Santos – nº 11997

Filipe Pereira Bastos – nº 13606

Orientadores:

Faculdade de Ciência e Tecnologia: Prof. Artur Miguel Dias

Novabase: Eng.º Pedro Moura

Monte da Caparica, Julho de 2004

Agradecimentos

Para a realização deste trabalho contamos com a preciosa ajuda de várias pessoas. Em primeiro lugar o nosso orientador, o prof. Artur Miguel Dias, que não sendo um especialista no assunto, se esforçou para nos ajudar dando, nalguns momentos, excelentes dicas de caminhos a seguir.

Um agradecimento muito especial vai também para o prof. Duarte Brito, prof. de Economia na FCT, que nos ajudou na escolha das variáveis relevantes economicamente e nos sugeriu características que o nosso sistema devia suportar.

À prof. Susana Nascimento e ao prof. Nuno Marques agradecemos a breve discussão que tivemos sobre Data Mining.

Na Novabase à que agradecer ao nosso orientador eng^o Pedro Moura que, apesar de ocupado com as suas obrigações profissionais, esteve disponível nos momentos mais importantes, nomeadamente ao indicar-nos o caminho relativamente às ferramentas de Data Mining a usar e a contactar outras pessoas na Novabase que nos pudessem também ajudar. Referimos o eng^o Humberto Neves que nos explicou uma metodologia de Data Mining que fora usada em projectos implementados pela Novabase nos seus clientes.

Ao Dr^o Jorge Cepeda agradecemos algumas noções teóricas sobre investimento na bolsa, nomeadamente o cálculo de suportes e resistências.

À Alexandra Pereira agradecemos o apoio sobretudo nos momentos finais deste projecto.

Resumo

O presente projecto tem como principal objectivo fornecer uma plataforma de apoio à decisão de investimento no mercado Bolsista. Com este intuito, desenvolveram-se ferramentas direccionadas para as seguintes áreas:

- Mecanismos automáticos de obtenção de dados via web.
- Mecanismos de visualização interactiva.
- Mecanismos de análise e interpretação de séries temporais (Indicadores) .
- Mecanismos de descoberta de padrões comportamentais (Candlesticks).
- Mecanismos inteligentes de descoberta de padrões (Data Mining).
- Mecanismos de simulação com base em dados passados.

O sistema de apoio à decisão construído é portanto bastante versátil e poderoso conjugando técnicas diferentes que pensamos constituírem uma mais valia importante para um investidor.

Índice

Agradecimentos	2
Resumo	3
Índice	4
Introdução	6
Contexto Académico.....	6
Contexto Tecnológico e Científico	6
Contexto Profissional.....	6
Objectivos do estágio.....	6
Estrutura do documento	6
Implementação	8
Recolha dos dados.....	8
Problemas na extracção de dados	8
Expressões regulares.....	8
Esquema da base de dados.....	9
Tabelas da Base de Dados.....	10
Indicadores.....	11
Implementação dos indicadores.....	15
Candlesticks	16
Visualização gráfica dos títulos e indicadores	19
Diagrama de Classes	21
Diagrama de Fluxo de Dados.....	21
Introdução ao Data Mining	22
Data Mining e Business Intelligence	23
Metodologia de Data Mining CRISP-DM	24
Data Mining na Bolsa de Valores	25
Prós e contras da aplicação de Data Mining na bolsa.....	25
Trabalho relacionado	25
Análise dos dados	27
Processo de Data Mining na bolsa de valores.....	31
Definição dos eventos de aprendizagem.....	33
Balanço do número de atributos com o número de eventos de treino	34
Overfitting.....	35
Criação da tabela de Data Mining.....	35
Escolha do algoritmo de classificação	35
Classificador C5.0.....	35
Treino de Classificadores.....	36
Automatização da geração de modelos no Clementine®	40
Avaliação dos modelos gerados pelo Clementine®	42
Análise Global para o conjunto de modelos do NASDAQ.....	42
Descrição da Aplicação de suporte à decisão	46
Módulo de Simulação	46
Módulo de análise técnica.....	48

Diferenças entre visualizar com candlesticks e com fecho ajustado	49
Módulo de análise da precisão das previsões dos modelos de Data Mining	50
Módulo de análise técnica.....	51
Conclusão	52
Apreciação Crítica do Trabalho Desenvolvido.....	52
Conhecimentos adquiridos.....	52
Trabalho Futuro	53
Aplicabilidade dos conhecimentos adquiridos na LEI.....	53
Apreciação do Estágio	53
Bibliografia.....	55
Anexo 1 - Descrição dos ficheiros de inicialização da base de dados.....	56
Anexo 2 - Dicionário de dados	59
Anexo 3 – Figuras para Análise Gráfica.....	64
Anexo 4 – Análise dos resultados dos modelos de Data Mining.....	67
10 Melhores Modelos	68
Anexo 5 – Ficheiros diversos.....	70
Ficheiros utilizados no processo de avaliação dos modelos de Data Mining	70
Ficheiros de Stream do Clementine®	70
Ficheiros para apoio na criação da tabela da Data Mining	70
Ficheiros de criação da base de Dados	70

Introdução

Contexto Académico

Este trabalho surge no âmbito do projecto de fim de curso da licenciatura de Engenharia Informática da Universidade Nova de Lisboa, Faculdade de Ciências e Tecnologia.

Contexto Tecnológico e Científico

As tecnologias usadas no presente estágio são as seguintes:

- Uso de Classificadores de Data Mining
- Bases de Dados
- Programação em C#

Contexto Profissional

É política da Novabase Business Intelligence (NBBI) que os estágios sirvam sobretudo como um período em que os alunos que lá estagiam se integrem na cultura da empresa e conheçam as ferramentas e técnicas com que, se o trabalho realizado for do agrado de ambas as partes, irão trabalhar depois do projecto terminar. Portanto a política não é a de estarmos em contacto com o cliente final a realizar um projecto para este dando assim lucro imediato à NBBI.

Objectivos do estágio

Ganhar competências na área do Data Mining e Business Intelligence e fazer a ligação entre o meio académico e empresarial.

Estrutura do documento

Implementação

Neste capítulo é descrita a forma de implementação da aplicação, desde a recolha de dados, passando pela criação da base de dados, são apresentados os diagramas de fluxos de dados e de classes de programação.

Introdução ao Data Mining

Neste capítulo é descrita a metodologia de Data Mining e a sua contribuição para o Business Intelligence.

Data Mining na Bolsa de Valores

Este capítulo descreve a aplicação de Data Mining à Bolsa de valores. Descreve as opções tomadas, as metodologias seguidas e os resultados obtidos.

Descrição da Aplicação de suporte à decisão

Neste capítulo é descrita a aplicação de suporte à decisão com as capacidades de visualização e análise implementadas. É aqui que é descrita a forma de utilização e funcionamento do ponto de vista do utilizador.

Conclusão

Por fim referem-se as conclusões a que se chegou ao levar a cabo este projecto. Conclusões ao nível do Data Mining sobre a bolsa, a funcionalidade de análise e simulação de estratégias de investimento, bem como possíveis melhorias a fazer numa futura continuação do projecto.

Implementação

Recolha dos dados

Uma fase crucial dum projecto de Data Mining é a recolha e qualidade dos dados. Os dados necessários são fornecidos por agências de informação como a Reuters. Felizmente há sites que têm a informação necessária. Depois de uma pequena pesquisa descobriu-se o site financeiro do Yahoo (<http://finance.yahoo.com>) que, além de ter toda a informação necessária até está, em geral, num formato que facilita o processamento automático.

A primeira fase do projecto foi então recolher os dados. Fez-se uma aplicação que consulta um url (retornando uma string gigantesca com o conteúdo) e depois processa-se essa string fazendo parsing ao seu conteúdo e introduzindo o que interessa nas tabelas da base de dados.

No anexo **Descrição das fontes de dados** descreve-se exaustivamente os endereços usados e os parâmetros necessários no url. Essencialmente consultamos 2 sites: o do Yahoo e o do Banco de Portugal. No site do Yahoo está a informação relativa a cada empresa (código empresa, industria, e histórico de cotações, dividendos e splits) e a cada índice (código do índice, composição, e histórico de cotações). Ao Banco de Portugal foi-se buscar informação diária relativa ao preço do ouro e das relações entre o euro e outras moedas.

Problemas na extracção de dados

No site do Yahoo, que é onde está o maior volume de dados, rapidamente nos apercebemos do mecanismo existente para evitar ataques Denial of Service. Depois de tirarmos consecutivamente dados do mesmo servidor já não se conseguia tirar mais.

Felizmente o Yahoo usa um esquema de replicação de dados que permite distribuir a carga. Como parte do url está o número de servidor. Gerando aleatoriamente este número resolveu-se o problema (a outra alternativa seria forçar alguns delays). Exemplo:

```
int randomBase = new Random().Next(3, 14);
//Yahoo servers are from 3 to 13, this way we avoid flooding the same server
string baseUrl = "http://ichart"+randomBase+".finance.dcn.yahoo.com/table.csv?";
```

Expressões regulares

De notar que, como seria de esperar, o url que contém os dados contém muita outra informação que não interessa para a nossa aplicação. Nesses casos foram necessárias expressões regulares. A class usada foi a RegEx do C#. Fica aqui um simples exemplo (contents é a string url completa):

```
private static string GetEmployees(string contents){
    string query =@">((\d+,\d\d\d)|(\d{1,3}))<";
    MatchCollection mc = Regex.Matches(contents, query, RegexOptions.Singleline);
    return mc.Count>0 ? Regex.Replace(mc[0].Groups[1].Value, ",", "") : "NULL";
}
```


Tabelas da Base de Dados

A informação retirada da fonte tem de ser devidamente organizada em tabelas para poder ser útil.

Fez-se 21 tabelas. Para maior detalhe nos campos consultar **Anexo Dicionário de Dados**.

As tabelas mais relevantes são:

Companies

Nesta tabela guarda-se informação relativa a uma empresa.

Dividendos

Esta tabela guarda a informação dos dividendos de cada companhia.

Um dividendo ocorre tipicamente ao fim de um ano (mas pode não ocorrer ou ocorrer mais frequentemente) e é quando a empresa partilha os lucros pelos seus accionistas.

Splits

Esta tabela guarda informação de split e mergers de empresas. Qual a data, qual a empresa e qual o racio de split (X acções antes do merger/split dão direito a Y acções depois).

Time

O tempo é um factor crucial na bolsa. Criou-se uma tabela que decompõe uma data nas várias componentes: dia do ano, dia do mês, ano, mês, semestre, trimestre, dia da semana, semana do ano, Esta informação é útil para pôr na tabela de Data Mining e pode ajudar o algoritmo de previsão.

Holidays

Crê-se que os feriados afectam de forma significativa o título nos dias próximos. Quando há um feriado não há negociação e isso pode ser um risco para os investidores. Pois assim ficam mais tempo sem puder movimentar o dinheiro se for caso disso. Nos dias que precedem um feriado pode haver uma ligeira pressão vendedora ou uma falta de iniciativa compradora.

Como alguns feriados são móveis e variam de país para país a técnica que adoptámos para os detectar foi, precisamente, descobrir os dias úteis em que não houve transacção. Até pode não ter sido um feriado, mas um dia em que a bolsa não abriu por motivos maiores, eg: atentados.

Os feriados estão associados a um país. Para determinar os feriados de um país, vamos ver quais os dias úteis em que não houve transacção em nenhum dos índices associados a esse país.

Stock quotes

É a maior tabela. Tem toda a informação relativa ao histórico de cada empresa.

Index quotes

Tem informação relativa ao histórico de cada índice.

Data Mining Table stock

Tem a informação toda desnormalizada relativa aos títulos que se querem analisar.

Indicadores

Objectivo dos indicadores

O grande volume de dados corresponde às cotações diárias na bolsa para cada título, ou seja, são séries temporais de valores. Para uma análise com algoritmos de aprendizagem automática com o objectivo de encontrar padrões na evolução destes valores é necessário considerar variáveis que definam para cada dia a informação relativa à evolução das cotações nos dias anteriores. Ou seja, são necessários indicadores que processem e agreguem a informação do passado.

Para extrair o conhecimento das séries de dados temporais foram usados indicadores de vários tipos, alguns dos quais largamente utilizados para análise técnica no mercado bolsista, outros foram estudados e criados para efeito deste projecto.

Os indicadores foram utilizados tanto para construir as tabelas de Data Mining como também para serem usados na interface gráfica incluída na aplicação, permitindo o seu acrescento ao gráfico de evolução diária das cotações. Desta forma, o formato dos valores calculados foi em muitos casos adaptado para servir da melhor forma ambas as aplicações.

Para que a informação contida nos indicadores seja bem utilizada pelos algoritmos de aprendizagem automática considerou-se que a melhor forma seria a utilização de valores facilmente comparáveis tais como rácios que definem as distancias a valores médios ou a linhas de suporte e resistência.

A forma como a informação agregada pelos indicadores é expressa, é essencial para que sua correcta captura na fase de aprendizagem dos classificadores.

Médias Móveis

As médias móveis permitem determinar divergências às tendências de curto ou de longo prazo. Foram calculadas para cada título médias móveis com intervalos variáveis (5, 20, 50 e 200 dias). Foram calculados três tipos de médias móveis, nomeadamente, médias móveis simples, pesadas e exponenciais.

As médias móveis simples (MA) limitam-se à soma de todos os elementos da janela, divididos pelo número de elementos.

As médias móveis pesadas (WMA) permitem atribuir maior peso aos valores mais recentes reagindo mais rapidamente às mudanças.

As médias móveis exponenciais (EMA) são calculadas afectando a média do dia anterior e o valor actual com pesos que permitem, tal como para as médias pesadas, atribuir maior peso aos valores mais recentes.

Para cada média móvel foram geradas 2 séries temporais, Uma com o valor da média e outra com a diferença do valor da cotação ao valor da média. Do ponto de vista do suporte à decisão usando a componente gráfica da aplicação o valor mais interessante é a média em si. No entanto, considerando a utilização algoritmos de aprendizagem automática, o melhor indicador será a diferença à média, pois não se tratam de valores absolutos e o treino do modelo mais dificilmente cai em situações de demasiada adequação aos dados do período de treino.

Indicador de força relativa

O indicador de força relativa (RSI) é um indicador de momento de oscilação muito popular, desenvolvido por J. Welles Wilder Jr (www.incrediblecharts.com).

O RSI é calculado considerando balanço dos movimentos descendentes e ascendentes numa determinada janela temporal. Para o presente projecto o valor dos movimentos descendentes e ascendentes foram calculados considerando os valores de abertura e fecho diários.

Formula de cálculo:

Média dos movimentos ascendentes = EMA(movimentos ascendentes)

Média dos movimentos descendentes = EMA(movimentos descendentes)

RS = Média dos movimentos ascendentes/ Média dos movimentos descendentes

RSI = $100 - 100 / (1 + RS)$

O valor do RSI é normalizado para o intervalo de 0 a 100. É frequentemente aceite que um valor abaixo de 30 corresponde a um sinal de compra, enquanto que um valor acima de 70 representa um sinal de venda. No entanto é comum variar estes valores dependendo da experiência de cada investidor.

Para o RSI foi gerada uma única série consistindo no valor do indicador.

Indicador convergência e divergência de médias móveis (MACD)

O indicador MACD consiste na medida da distância entre duas médias móveis com janelas temporais diferentes (geralmente 12 e 26 dias). Uma linha de sinal é calculada com a média exponencial de 9 dias do valor do indicador MACD. Quando a linha do MACD cruza a linha de sinal surgem os sinais de compra ou venda.

Sinal de venda – o MACD cruza a linha de sinal vindo de baixo para cima

Sinal de compra – o MACD cruza a linha de sinal vindo de cima para baixo

Para o MACD foram geradas duas séries, respectivamente a série do MACD e a série do sinal. Ao tomar esta opção considerou-se que posteriormente os classificadores a utilizar teriam a capacidade de relacionar estas duas grandezas caso de facto fossem determinantes para determinar a evolução das cotações na bolsa.

Indicador Estocástico

O indicador estocástico foi desenvolvido por Dr. George Lane para monitorizar o momentum do mercado bolsista. Existem duas versões do indicador, o *fast Stochastic* e o *slow Stochastic*. Para ambas as versões o indicador consiste em duas séries de valores:

- %K faz um balanço entre as amplitudes entre valores máximos e mínimos para um intervalo e a amplitude entre o preço de fecho mínimo e o de hoje.
- %D é o resultado de uma suavização da linha %K com uma média móvel

Método de calculo:

CL = fecho Hoje – valor mínimo no intervalo

HL = valor máximo no intervalo – valor mínimo no intervalo

$\%K = CL/HL * 100$

$\%D = MA(\%K)$ a 3 dias

Para o *slow Stochastic* a linha %K é igual é linha %D do *fast Stochastic* e a linha %D corresponde á média móvel da linha %K.

Para este indicador foram geradas 3 séries, respectivamente %K fastStoch, %D fastStoch e %D slowStoch. A série %K slowStoch foi ignorada já que é equivalente ao %D fastStoch.

Indicador de rácio de mudança (ROC)

Este indicador foi desenhado para detectar pontos de inversão de tendências. A forma de cálculo é a seguinte:

$$\text{ROC} = ((\text{Valor de fecho a N dias atrás} - \text{Valor de fecho hoje}) / \text{Valor de fecho N dias atrás}) * 100$$

- Quando o indicador passa de um valor negativo para positivo indica tendência de subida
- Quando o indicador passa de um valor positivo para negativo indica tendência de descida

Para este indicador foi calculada uma série única com o valor do ROC.

Suportes Resistências

O suporte corresponde ao valor a partir do qual se prevê que a cotação de um título tenha tendência a inverter movimentos descendentes. A resistência corresponde ao conceito simétrico do suporte, é aquele valor a partir do qual se espera que a tendência de subida de um título seja quebrada.

Estes limites podem ser justificados empiricamente pelo facto de a grande maioria do mercado ser levada a comprar quando as cotações atingem determinados valores que são considerados subvalorizados sendo que a forte procura dos títulos nessas alturas leva á quebra da tendência de descida, ou inversamente, a venda quando o título é considerado sobrevalorizado leva a posterior quebra do movimento ascendente.

Os valores de resistência e suporte devem ser periodicamente ajustados considerando a evolução da cotação.

Os suportes e resistências foram calculados tendo em conta os valores máximos e mínimos para um dado título num dado intervalo de tempo. Considerando estes valores, verifica-se frequentemente que quando um máximo é atingido e se inicia a descida da cotação, esta, terá tendência a inverter essencialmente em 3 pontos definidos entre o máximo e o mínimo para o intervalo desde que o valor mínimo ou máximo não sejam quebrados. Estes pontos foram calculados recorrendo a rácios extraídos da série de fibonnaci. São respectivamente 0,375, 0,5 e 0,625. Esta técnica foi testada e afinada com os dados que possuímos e revelou-se de forma geral relativamente precisa embora não existe uma explicação teórica fundamentada para esta técnica.

Quando os mínimos e máximos que permitem determinar um dado suporte e resistência são ultrapassados a resistência antiga passa a ser o suporte ou vice versa.

O resultado cálculo dos suportes e resistências foi implementado de forma a permitir a sua interpretação por um algoritmos de aprendizagem automática. Assim foram geradas 6 séries de valores, respectivamente, o valor do suporte, a diferença em percentagem da cotação ao valor do

suporte, o número de vezes que o suporte foi tocado pela cotação e as mesmas grandezas para a resistência.

Evolução passada

Considerou-se que um indicador que poderia ser importante para treino de modelos de Data Mining seria a evolução nos últimos n dias até à data presente pois a evolução de um título pode ser fortemente condicionada mediante a sua evolução nos últimos dias. Assim procedeu-se ao cálculo destes valores para intervalos temporais semelhantes aos das variáveis objectivo mas no sentido do passado.

Variáveis objectivo

Para determinar as variáveis objectivo (ver capítulo de Data Mining na Bolsa de Valores) bastou para cada série de dados verificar a evolução em percentagem em relação a cada uma das datas da série.

Implementação dos indicadores

A aplicação desenvolvida inclui três classes responsáveis pela gestão dos indicadores. O repositório de um Indicador (Indicator), a classe responsável pela manipulação e inserção na base de dados (IndicatorCalculation) e a classe com a definição das funções dos indicadores (IndicatorFunctions).

Os indicadores foram calculados e inseridos na base de dados para cada título, índice e indicadores mundiais. Considerou-se que, do ponto de vista do Data Mining, seria esta a melhor forma de analisar não só a informação das cotações como também a informação dos índices e dos indicadores mundiais. Cada um destes conjuntos de indicadores foi inserido em tabelas independentes.

A inserção na base de dados é efectuada individualmente para cada ticker (código do título, índice ou indicador mundial). Em primeiro lugar é verificado na base de dados qual a data para a qual já existem indicadores calculados e em seguida, considerando os nomes das colunas da tabela a preencher são determinados quais os indicadores a calcular. Desta forma, para alterar o tipo de indicadores a calcular basta alterar a tabela para que o nome seja identificado como um novo indicador e com os respectivos parâmetros de configuração.

A classe IndicatorFunctions é a responsável pela interpretação dos nomes dos indicadores e dos respectivos parâmetros, No início do preenchimento da base de dados são passados para esta classe os nomes dos indicadores sendo que para os indicadores reconhecidos são gerados objectos indicator a partir das funções respectivas de cálculo.

A classe IndicatorCalculation permite a funcionalidade de actualização com os valores de indicadores para as novas cotações obtidas periodicamente a partir da Internet.

Candlesticks

A informação relativa a um dia (sessão) é o preço de abertura, fecho, máximo e mínimo. Essa informação consegue ser condensada numa única figura chamada Candlestick.

Se o candlestick for verde significa que nesse dia houve uma subida: o valor de fecho é superior ao da abertura.

Alguns livros representam este candlestick a branco.

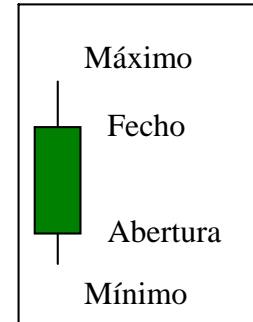


Figura 1

Se o candlestick for vermelho significa que nesse dia houve uma descida: o valor de fecho é inferior ao da abertura.

Alguns livros representam este candlestick a preto.

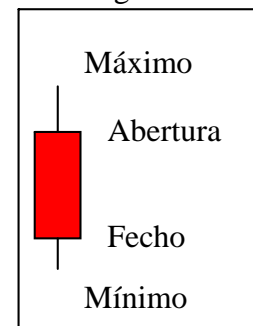


Figura 2

Consoante a imagem formada o candlestick tem um nome próprio. Eis alguns exemplos:

Long Black Line



Long White Line



Dragonfly Doji



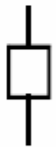
Four-Price Doji



Black Spinning Top



White Spinning Top



Small Black Doji



Small White Doji



Long Black Marubozu



Long White Marubozu



Black Umbrella



White Umbrella



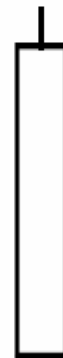
Hammer or Hangman

Hammer or Hangman

Long Black Closing Bozu



Long White Closing Bozu



Gravestone Doji



Long-Legged Shadow Doji



Gravestone Doji

Long-Legged Shadow Doji

Para uma lista exaustiva deste candlesticks consultar referência [3].

Quando um determinado conjunto de candlesticks aparece forma-se um padrão. Há uma lista de padrões referenciados que têm um significado associado (subida ou queda).

O padrão verifica-se no dia em que ocorre o candlestick de confirmação.

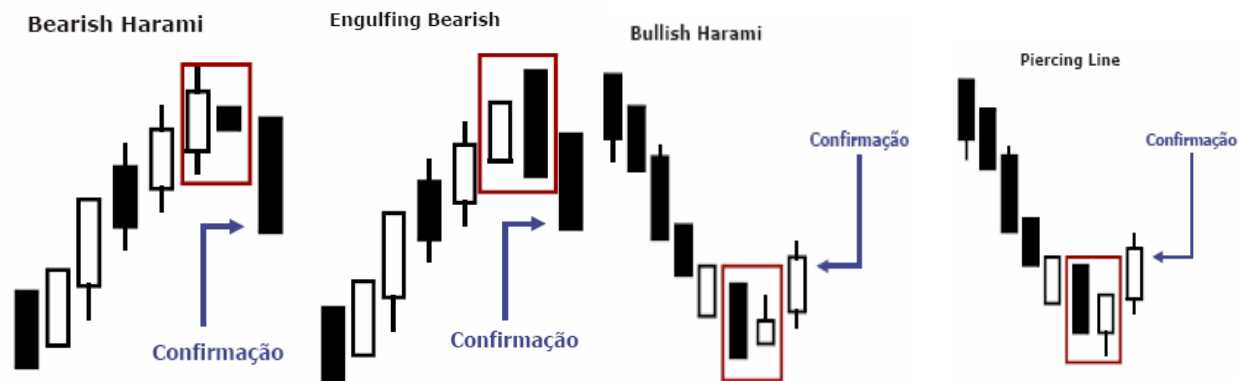


Figura 3: Exemplos de padrões teóricos de candlestick

Na nossa ferramenta de apoio à decisão implementámos um algoritmo de detecção automática de padrões.

Detecção de figuras isoladas

Criámos uma classe para identificar figuras isoladamente.

Essa classe recebe 4 números (abertura, máximo, mínimo e fecho) e, a partir destes, tenta inferir de que figura se trata.

Para cada uma das figuras teóricas sabemos qual a relação exacta entre a amplitude (high-low), a abertura e o fecho.

Dificilmente os rácios calculados para os números que estamos a avaliar encaixam numa das figuras teóricas. Há sempre um erro associado. Dizemos que a figura teórica associada é a que minimiza o erro, desde que este esteja abaixo de um épsilon por nós escolhido.

Caso não seja possível detectar fica como unknown figure

Detecção de padrões

A classe que identifica padrões baseia-se na identificação de figuras.

Para simplificar considera-se que um padrão é composto exactamente por 8 figuras simples, o que em geral é verdade.

O mais importante na detecção de um padrão é a formação do monte (ou vale) e as últimas 3 figuras.

É importante que o candlestick de confirmação seja do mesmo tipo do teórico assim como a última e penúltima figura.

Exemplo de detecção



Figura 4: Exemplo de detecção de padrão de candlestick pelo nosso sistema

O nosso sistema detecta a confirmação do candlestick Engulfing Bullish dia 21 de Junho. (seria portanto uma boa altura para comprar segundo [4])

Visualização gráfica dos títulos e indicadores

De forma a tornar mais fácil a percepção dos dados decidimos mostrar graficamente informação relativa a uma empresa. Além de mostrarmos os dividendos, merger e splits e informação diversa sobre a empresa (toda disponível na nossa base de dados), mostramos também o valor de várias séries.

O gráfico pode ser de 2 tipos: Candlesticks (onde cada candlestick mostra a informação relativa a um dia: open, high, low e close: ver secção Detecção de padrões) ou simples onde só mostramos o valor de fecho ajustado. (o valor de fecho ajustado já desconta dividendos e mergers ou splits)

Além da série do título permitimos também a escolha de indicadores que serão mostrados. Esses indicadores são preciosos na análise técnica. Ligando a opção “série details” aparecem na memo box respectiva os valores exactos dos indicadores seleccionados para o dia corrente.



Figura 5: Mostrando as potencialidades do nosso ambiente de análise integrado

Diagrama de Classes

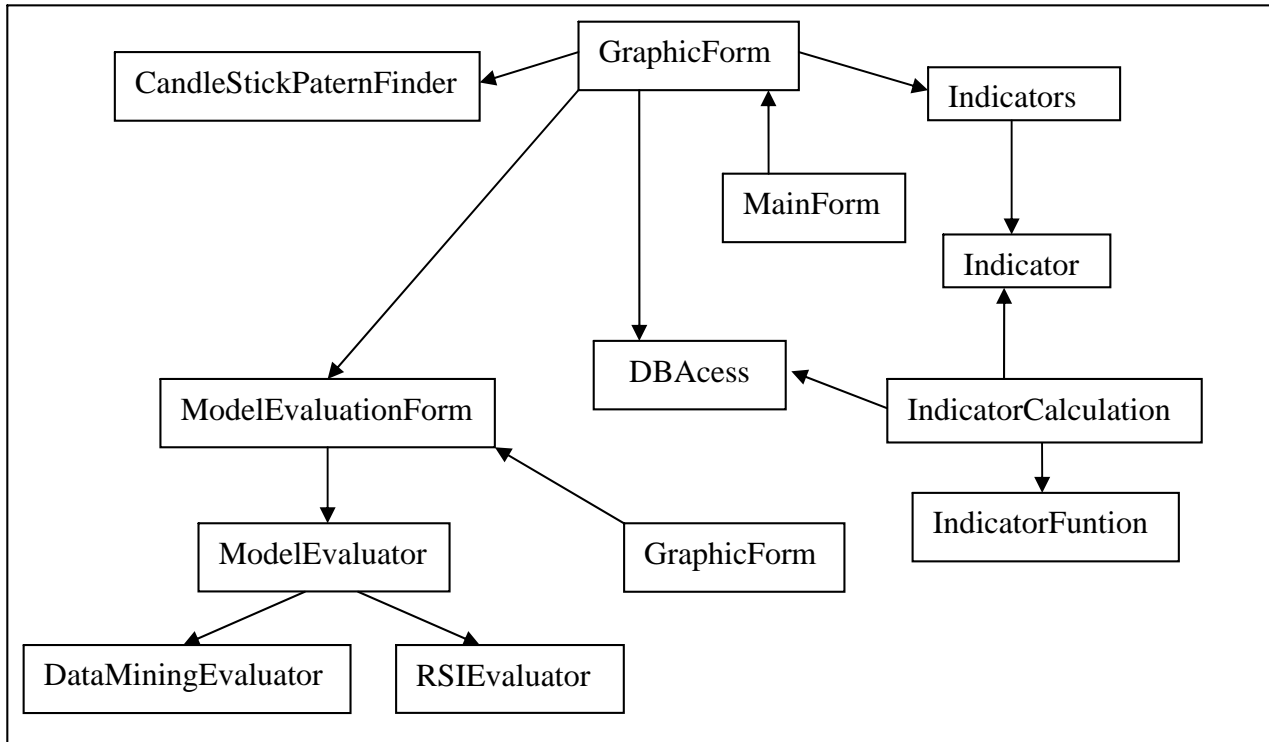


Figura 6: Diagrama de classes

Diagrama de Fluxo de Dados

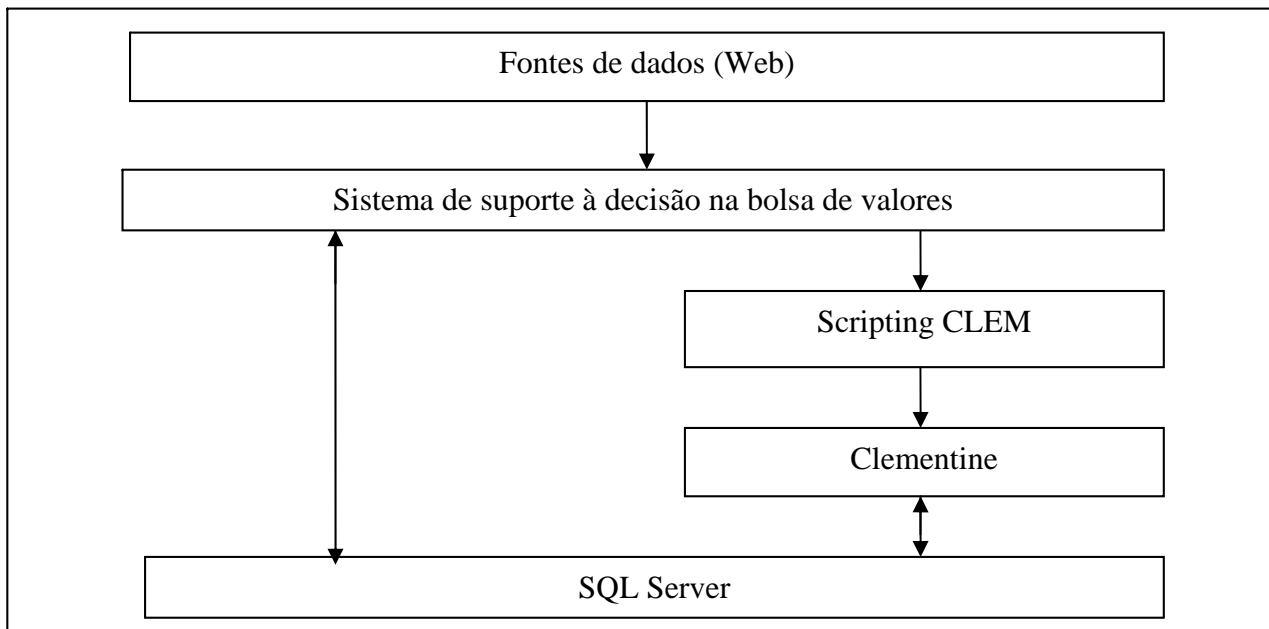


Figura 7: Diagrama dos fluxos de informação entre os principais módulos.

Introdução ao Data Mining

Data Mining é o processo de extrair informação válida previamente desconhecida a partir de grandes bases de dados e em seguida utilizar essa informação como suporte à decisão dos gestores do negócio. O Data Mining é uma área interdisciplinar que junta técnicas como a aprendizagem automática, reconhecimento de padrões, estatística, bases de dados e técnicas de visualização de dados.

Motivação

Os motivos que têm levado à crescente utilização de Data Mining, pertencem essencialmente aos seguintes grupos:

- **Relação com o cliente:**

A descoberta de conhecimento, por Data Mining, pode levar a um rejuvenescimento da relação com o cliente. Os serviços fornecidos podem ser mais personalizados, através do estudo dos padrões dos clientes tornando mais próxima a relação empresa-cliente.

- **Competitivos entre empresas**

Com o aumento da competitividade entre empresas, o ganho de conhecimento por tecnologias de Data Mining pode levar a uma mais valia na previsão de cenários e à antecipação a acontecimentos futuros. Descoberta de grupos de clientes mais susceptíveis a um determinado produto.

- **Combate à fraude**

A descoberta dos padrões que estão por trás de comportamentos fraudulentos podem levar a maior eficiência nas acções de fiscalização ao reduzir o universo de pesquisa necessário para encontrar um maior número de casos.

Suporte tecnológico

Cada vez mais em qualquer actividade económica os dados são guardados em grandes volumes. Em cada 20 meses o volume de dados mundial guardados em bases de dados duplica de volume [9]. Esta enorme quantidade de dados guarda em si intrinsecamente muito conhecimento que sem técnicas adequadas fica dissimulado. Desde sempre o homem procurou interpretar e extrair conhecimento de volumes de dados, mas nunca antes foi tão fácil o acesso à informação e nunca antes esteve presente uma tão grande capacidade de processamento de dados.

Estes são os factores que potenciam a utilização da Data Mining e colocam este tipo de tecnologias no topo das ferramentas de suporte à decisão para apoio à gestão.

Os dois principais objectivos em Data Mining tendem a ser a predição e a descrição de dados. A predição de dados envolve o uso de um conjunto de variáveis para prever os valores desconhecidos ou futuros de uma outra variável objectivo. A descrição de dados consiste na descoberta de padrões que descrevam os dados. No presente trabalho pretende-se o treino de classificadores no sentido de prever valores futuros com base no comportamento passado.

Data Mining e Business Intelligence

O Data Mining actua em parceria com outras técnicas de análise de dados de primeira linha, como o *data warehousing*, *reporting*, *online analytical processing* (OLAP) e análise estatística. As restantes técnicas de Business Intelligence permitem a manipulação da informação mas não permitem por si só a descoberta de informação. É aqui que entram as potencialidades do Data Mining, as ferramentas de Data Mining permitem descobrir conhecimento anteriormente escondido na massa de dados bem como a confirmação de regras de negócio já anteriormente conhecidas.

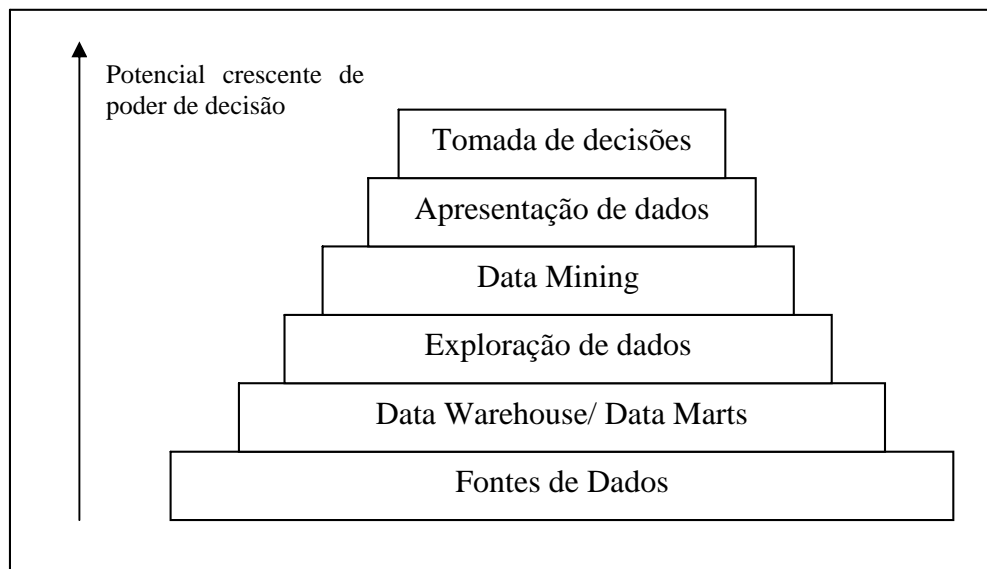


Figura 8: Posição do Data Mining na cadeia de ferramentas de Business Intelligence (Fonte: Discovering Data Mining From Concept to Implementation)

Metodologia de Data Mining CRISP-DM

A metodologia standard CRISP-DM surgiu em 1996 criada por três veteranos da aplicação de Data Mining a problemas reais. Até 1999 muitos progressos foram feitos para chegar à metodologia CRISP-DM 1.0. Esta metodologia foi criada com base na experiência de aplicação de Data Mining a problemas do mundo real e como tal é muito orientada para os aspectos práticos dessa problemática.

Na sua essência o método CRISP-DM 1.0 é descrito como um processo hierárquico no qual se podem identificar 4 níveis principais.

- 1) Conjunto de fases do processo de Data Mining.
- 2) A cada fase corresponde um conjunto de tarefas genéricas.
- 3) A cada conjunto de tarefas genéricas é associado um conjunto de tarefas especializadas.
- 4) Registo das acções e decisões do processo de Data Mining de acordo com as tarefas definidas nos níveis superiores.

As principais fases consideradas no processo de Data Mining são:

- Compreensão do Negócio
- Compreensão dos Dados
- Preparação dos dados
- Modelação
- Avaliação
- Aplicação

A figura seguinte (Figura 4) mostra o ciclo de um processo de Data Mining considerando as várias fases já enumeradas. O processo de Data Mining pode ser encarado como um ciclo com continuo aperfeiçoamento pelo aproveitamento da experiência acumulada e pela análise de novos dados quer de negócio quer dos próprios resultados da aplicação dos modelos.

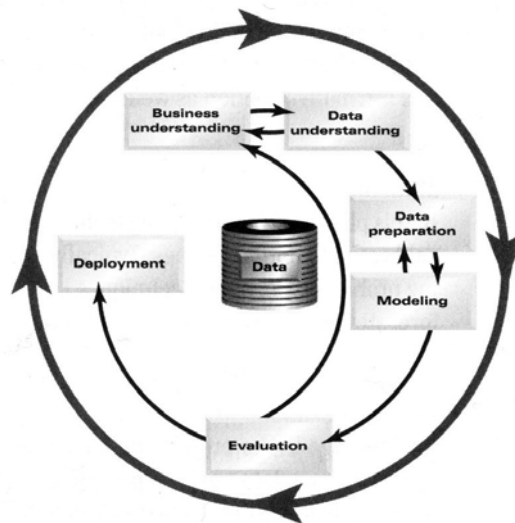


Figura 9: Fases da metodologia CRISP-DM 1.0

Data Mining na Bolsa de Valores

Prós e contras da aplicação de Data Mining na bolsa

O comportamento da bolsa é orientado por pressupostos altamente voláteis, logo, conhecer o comportamento no passado não é garantia para prever o comportamento futuro.

A evolução das cotações pode depender de factores dificilmente interpretáveis por algoritmos de inteligência artificial, tais como notícias com repercussões negativas.

É impossível agregar a quantidade de informação necessária para representar o conjunto de variáveis que justificam a evolução da bolsa, até porque muitas vezes segue padrões comportamentais que desafiam a lógica de mercado.

Os indicadores da bolsa, por vezes surpreendentemente, até dão bons resultados, no entanto é difícil escolher a melhor interpretação conjugada dos mesmos para definir estratégias de investimento, as ferramentas de aprendizagem automática podem ajudar a conjugar toda esta informação da melhor forma.

Desde que seja possível reunir um volume de dados representativo, teoricamente seria possível extrair informação útil que usando ferramentas menos poderosas não seria descoberta.

Trabalho relacionado

A aplicação de Data Mining na tentativa de prever o comportamento da bolsa, tem sido largamente investigada nos últimos anos e tem dado origem aos mais variados resultados.

Com a disponibilização de grandes volumes de dados históricos, via web, a massificação de dados tornou possível a utilização de algoritmos de aprendizagem automática em larga escala. Muitos padrões no comportamento das cotações na bolsa foram descobertas utilizando ferramentas de Data Mining sobre as cotações passadas em conjunto com outras grandezas relacionadas (por vezes sem relação aparente).

A passagem de regras do comportamento da bolsa resultantes de Data Mining para estratégias de investimento enfrenta obstáculos de peso. É essencial ter noção se um determinado padrão comportamental descoberto por Data Mining deve-se apenas a um acaso nos dados de treino ou se existe fundamento estatístico para ser utilizado na predição da evolução futura.

Numa aplicação de Data Mining para previsão da evolução do índice S&P 500 feita por Leinweber verificou-se que o atributo com maior correlação era a produção de manteiga no Bangladesh (75% de correlação num período de 20 anos) [9]. Este tipo de relações pode acontecer sobretudo se o balanço entre o número de atributos e o número de registos não for adequado. Quando o número de atributos é excessivo facilmente, podem ocorrer correlações casuais entre os dados.

Naturalmente, os padrões de comportamento descobertos utilizando ferramentas de Data Mining são considerados estatisticamente mais representativos quando maior for o período de pesquisa. No entanto, depois de descoberto e divulgado um padrão que representa uma oportunidade de investimento, este tende a deixar de existir no futuro à medida que um número crescente de investidores o tenta aproveitar influenciando o evoluir das cotações.

Análise dos dados

Na fase inicial de qualquer trabalho de Data Mining é essencial o conhecimento a fundo do universo de dados com o qual vamos lidar. Para isso devemos proceder a um conjunto de análises para entender as possíveis relações que se podem observar usando ferramentas de análise tradicionais. É deste tipo de análise que pode ser estabelecida a estratégia de triagem e processamento da informação de forma a apresentar aquela mais relevante e no melhor formato, ao algoritmo de aprendizagem automática. Para esta análise usámos as cotações dos títulos pertencentes ao índice NASDAQ entre os anos de 1997 e 2003 inclusivamente. Para este período dispomos de um total de 167.954 registos de cotações diárias.

Relação entre a evolução das cotações durante 1 dia e o dia da semana

Uma das questões de interesse seria verificar se a evolução das cotações na bolsa diferem em média consoante o dia da semana. Observando o gráfico da figura 1 podemos concluir que para o período de 1997 a 2003 o pior dia de bolsa tem sido a segunda-feira. Enquanto que o melhor dia de bolsa é em média a quinta-feira. Considerando esta diferença diária consideramos que o dia da semana poderá ser um atributos valioso para treino de modelos de Data Mining.

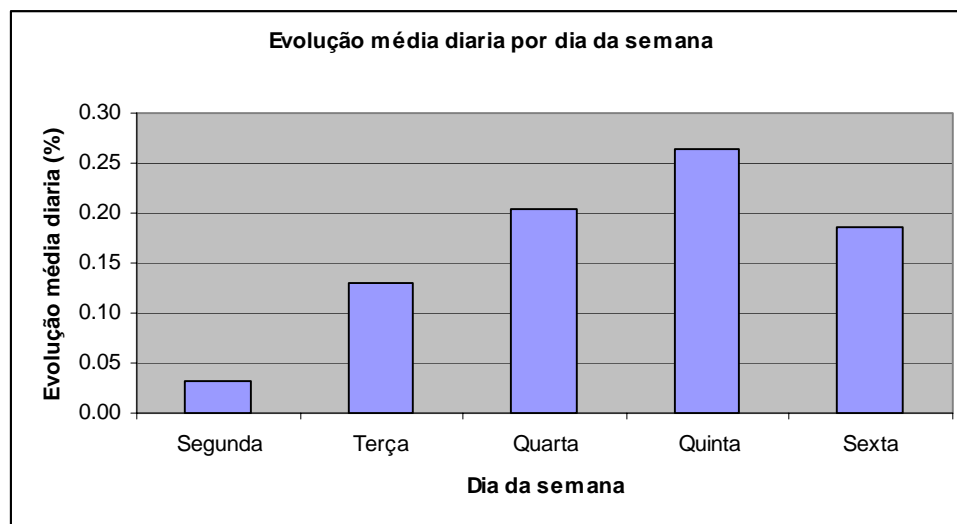


Figura 10: evolução das cotações diárias para os títulos do índice NASDAQ para cada dia da semana

Evolução da cotação nos dias que antecedem uma distribuição de dividendos

Seria de esperar que quando se aproxima uma distribuição de dividendos e logo após a mesma o comportamento das evoluções diárias das cotações fosse de alguma forma afectado. Considerando a figura seguinte podemos verificar que em média observa-se uma desvalorização acentuada nos primeiros 4 dias, recuperada ao 5º dia.

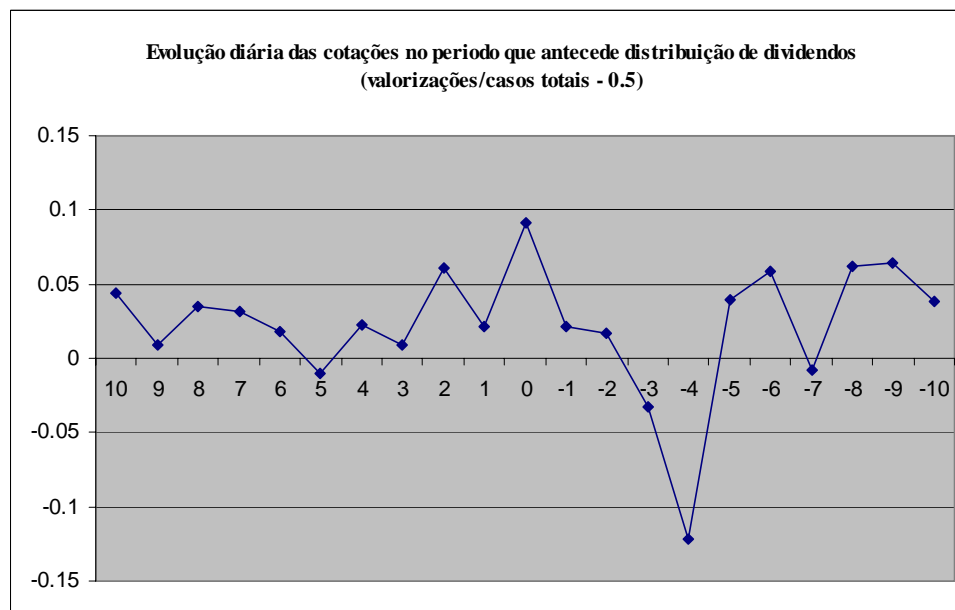


Figura 11: Evolução diária das cotações para os títulos do índice NASDAQ na proximidade das distribuições de dividendos

Evolução da cotação após grandes valorizações

Com o objectivo de verificar se após grandes valorizações existem comportamentos distintos para a maioria dos títulos, escolheram-se aqueles casos cuja cotação subiu mais do que 15% em 1 dia, tendo-se obtido os resultados do quadro seguinte:

Quadro 1: distribuição da evolução em duas classes (valorização e desvalorização) quando no dia anterior aconteceu uma valorização superior a 15%

	Evolução	Proporção	%	Número de casos
Evolução a 1 dia	desvalorização		45,57	432
	valorização		54,43	516
Evolução a 3 dias	desvalorização		47,57	451
	valorização		52,43	497
Evolução a 7 dias	desvalorização		51,58	489
	valorização		48,42	459
Evolução a 15 dias	desvalorização		59,60	565
	valorização		40,40	383









Como se pode verificar pelo quadro anterior, logo após uma valorização forte a tendência é de queda no primeiro dia, mas nos dias seguintes a tendência para a valorização sobrepõe-se chegando a 59% de valorização 15 dias após a forte subida.

Considerando a sua aparente relação com o comportamento da bolsa, a evolução nos dias anteriores deve ser considerada como um indicador valioso.

Evolução da cotação após grandes desvalorizações

Nos períodos que se seguem a grandes desvalorizações (>15% em 1 dia) observou-se o comportamento patente no quadro seguinte:

Quadro 2: distribuição da evolução em duas classes (valorização e desvalorização) quando no dia anterior aconteceu uma desvalorização superior a 15%







	Evolução	Proporção	%	Número de casos
Evolução a 1 dia	desvalorização		38,88	180
	valorização		61,12	283
Evolução a 3 dias	desvalorização		40,39	187
	valorização		59,61	276
Evolução a 7 dias	desvalorização		42,76	198
	valorização		57,24	265
Evolução a 15 dias	desvalorização		38,23	177
	valorização		61,77	286

Também as grandes desvalorizações mostram uma forte relação com o comportamento da bolsa nos dias seguintes. Após uma forte desvalorização a tendência é sempre de recuperação sendo mais evidente 15 dias após a ocorrência.

Evolução da cotação quando o RSI é superior ou igual a 80

Observando os histogramas para a evolução das cotações da bolsa quando o indicador RSI é superior a 80 esperamos encontrar um desvio representativo de uma maior tendência de queda nos dias seguintes.

Quadro 3: distribuição da evolução em duas classes (valorização e desvalorização) quando o RSI ≥ 80

	Evolução	Proporção	%	Número de casos
Evolução a 1 dia	desvalorização		52,16	3184
	valorização		47,84	2920
Evolução a 3 dias	desvalorização		52,46	3202
	valorização		47,54	2902
Evolução a 7 dias	desvalorização		49,69	3033
	valorização		50,31	3071

Considerando a análise gráfica verificamos que quando o indicador RSI se encontra acima de 80 nota-se uma ligeira tendência de descida da cotação, que é mais forte na evolução a 1 e 3 dias e que já não se faz sentir a 7 dias.

Evolução da cotação quando o RSI é inferior ou igual a 20

Observando os histogramas para a evolução das cotações da bolsa quando o indicador RSI é inferior a 20 esperamos encontrar um desvio representativo de uma maior tendência de subida nos dias seguintes.

Quadro 4: distribuição da evolução em duas classes (valorização e desvalorização) quando o RSI ≤ 20

	Evolução	Proporção	%	Número de casos
Evolução a 1 dia	desvalorização		46,01	2798
	valorização		53,99	3283
Evolução a 3 dias	desvalorização		43,89	2669
	valorização		56,11	3412
Evolução a 7 dias	desvalorização		42,58	2589
	valorização		57,42	3492

Considerando a análise gráfica podemos verificar que o indicador RSI acrescenta uma mais valia à previsão da evolução das cotações na bolsa. Por este motivo devemos incluir este indicador no treino de modelos de Data Mining.

Processo de Data Mining na bolsa de valores

O processo de Data Mining seguido no âmbito deste projecto está exemplificado na figura seguinte (Fig. 6):

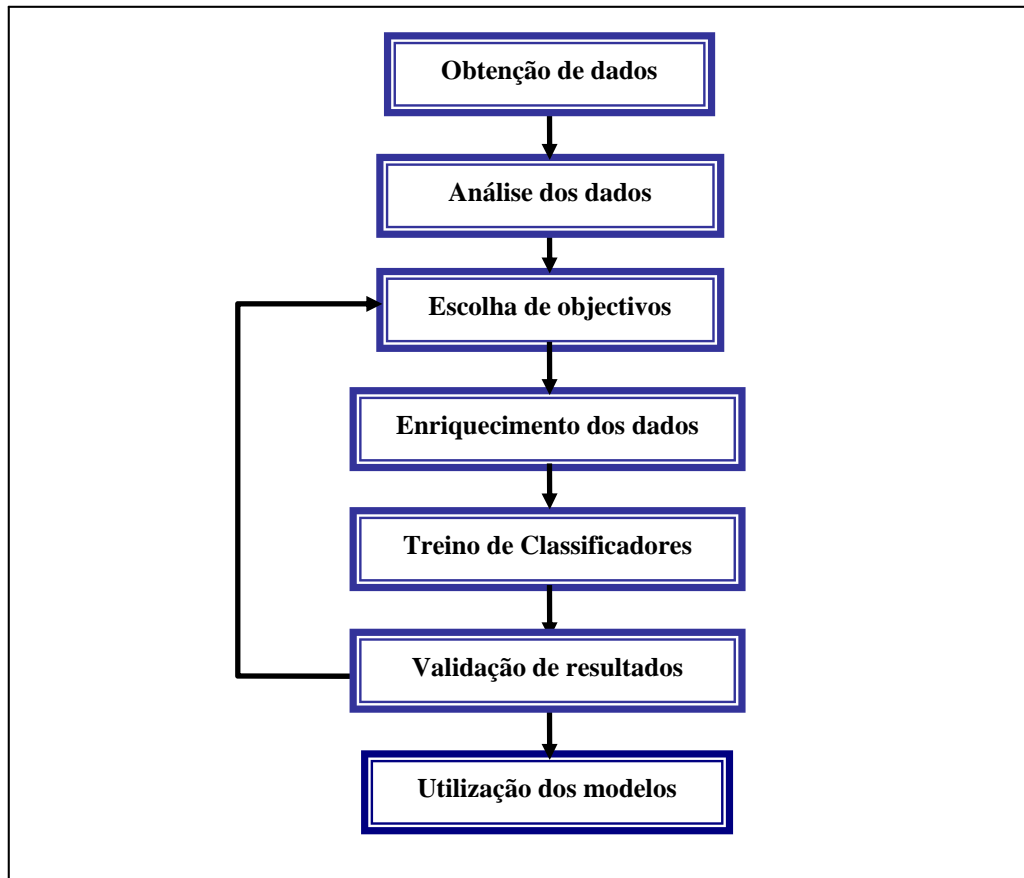


Figura 12: Esquema do processo de Data Mining

Obtenção de dados

Em primeiro lugar deve ser pesquisada a qualidade e quantidade de informação que pode ser obtida considerando quais as grandezas relevantes na tentativa de estimar o comportamento da bolsa de valores.

Análise dos dados

Para conhecer ou confirmar as regras de negócio é essencial proceder à análise dos dados. Esta fase dá suporte aos passos seguintes de definição de objectivos e enriquecimento de dados.

Definição de Objectivos

Em segundo lugar, para uma correcta aplicação de ferramentas de Data Mining é necessário entender o problema a modelar e quais os objectivos que podem ser propostos. Os objectivos devem ser ambiciosos tanto quanto possível de acordo com o problema mas considerando a qualidade e quantidade de dados disponíveis.

Considerando o mercado bolsista e os dados disponíveis determinamos que a previsão da percentagem de evolução futura seria o objectivo a perseguir.

A melhor forma de verificar a capacidade predictiva dos modelos teve que passar pela definição de diferentes intervalos temporais, já que seria de esperar que, consoante o intervalo, assim varie a capacidade predictiva dos modelos. Já que à priori não seria possível fazer uma estimativa dos melhores intervalos de tempo, optámos por percorrer o espaço temporal de dois meses escolhendo os valores que considerámos interessantes do ponto do vista de futuras estratégias de investimento de curto e médio prazo.

Os objectivos escolhidos foram a percentagem de evolução a 1, 3, 7, 15, 30 e 60 dias a partir da data actual.

Posteriormente para cada um destes intervalos foi possível verificar quais permitem uma maior capacidade predictiva para os respectivos modelos.

Para cada um destes intervalos, os valores de evolução futura foram discretizados em classes de forma a permitir a utilização de classificadores do tipo C5.0 (árvores de decisão) que requerem, que as variáveis objectivo sejam conjuntos finitos.

As classes escolhidas foram igualmente a título experimental, para posteriormente serem verificadas quais as melhores soluções de discretização.

Assim as discretizações escolhidas foram respectivamente:

2 Classes:

- evolução $\leq 0\%$
- evolução $> 0\%$

3 Classes:

- evolução $< -1\%$
- evolução entre -1% e 1%
- evolução $> 1\%$

5 Classes:

- evolução $< -3\%$
- evolução entre -3% e -1%

- evolução entre -1% e 1%
- evolução entre 1% e 3%
- evolução entre >3%

Desta forma, ao todo, foram definidos inicialmente 6 intervalos temporais, e para cada um deles 3 conjuntos de classes, totalizando 18 objectivos distintos para cada conjunto de dados de treino. A cada objectivo corresponderá o treino de um modelo distinto.

Quando maior o número de classes de discretização é de esperar uma maior dificuldade de classificação, mas do ponto de vista da utilização para previsão da evolução da bolsa, podem daí surgir vantagens na elaboração de estratégias de investimento.

Enriquecimento de dados

A fase de enriquecimento de dados, consistiu em todo o cálculo de variáveis com base nas séries temporais descritos no capítulo de implementação.

Definição dos critérios de sucesso

O critério de sucesso no que diz respeito à aplicação de Data Mining na bolsa de valores com os objectivos propostos, corresponde ao ganho de informação na previsão da evolução face à escolha aleatória. Ou seja, considerando os objectivos com discretização em duas classes, e tendo em conta que a quantidade de casos com evolução positiva e negativa são a longo prazo bastante equilibrados, o ganho de informação de um modelo de previsão corresponde a todo o acréscimo de precisão acima dos 50%.

Definição dos eventos de aprendizagem

O evento de aprendizagem consiste na menor unidade de informação usada para cada caso de treino do algoritmo de aprendizagem automática. Para o mercado bolsista e tendo em conta os objectivos propostos considerou-se que um evento de treino corresponde a um *snapshot* para uma data, dos atributos que caracterizam a situação do título e o respectivo valor para a variável objectivo devidamente discretizada. Para cada evento foi determinada a evolução da cotação nos dias seguintes à data do *snapshot* de acordo com os intervalos de tempo definidos nos objectivos (Figura 7).

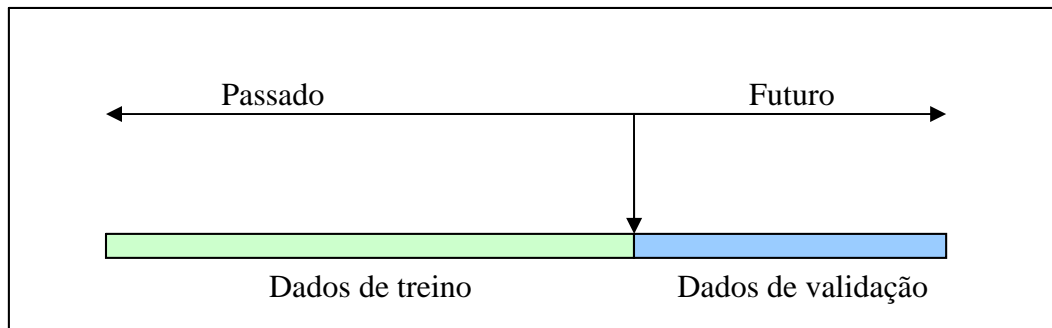


Figura 13: Definição dos eventos de treino

O formato da informação requerida para os algoritmos de aprendizagem automática é essencialmente uma tabela única em que cada linha contém um conjunto de condições e o valor da variável objectivo. No presente trabalho o valor das variáveis objectivo corresponde às evoluções futuras na cotação da bolsa para um dado título a partir de uma determinada data.

Um evento de treino é construído, definindo em primeiro lugar, uma data que representa o tempo presente, ou seja, a data do evento. Para esta data, e para seu passado relativo, é obtida toda a informação considerada relevante para definir o estado do mundo a modelar. O campo objectivo será o que na realidade aconteceu após essa data de acordo com os objectivos propostos. Assim sendo, as condições devem ser sempre definidas com coerência temporal, ou seja, nunca podem ser definidas com informação posterior à data do evento pois desta forma estaríamos a fornecer ao modelo informação sobre o futuro.

Balanco do número de atributos com o número de eventos de treino

Uma problemática de extrema relevância para o Data Mining é o balanço do número de atributos com o número de eventos de treino. Quando o número de atributos é demasiadamente elevado podem com maior probabilidade, surgir correlações que não passam de coincidências pontuais e que por sua vez podem levar a que os modelos gerados as interpretem como características intrínsecas dos dados.

Desta forma, os modelos podem perder generalidade e, conseqüentemente, capacidade de previsão quando utilizados com dados não incluídos no conjunto de treino.

Este problema foi largamente invocado quando da utilização de ferramentas de Data Mining sobre a bolsa utilizando um excessivo número de atributos.

Os atributos devem ser escolhidos considerando que as relações entre eles são prováveis e que podem fazer sentido segundo as regras de funcionamento do problema a modelar.

Overfitting

Quando se usam algoritmos de classificação para posterior previsão deve ter-se em consideração que, caso a aprendizagem seja demasiado exaustiva sobre os dados usados pode construir-se um modelo demasiado adequado para os dados de treino sendo posteriormente pouco adaptável a dados futuros não incluídos no grupo de treino.

Quando da definição dos parâmetros dos classificadores deve procurar criar-se um modelo generalista de forma a ter maior capacidade de adaptação aos dados não incluídos no conjunto de treino. Consoante o algoritmo de aprendizagem usado e o tipo de dados é possível definir uma metodologia para evitar o overfitting do modelo.

Criação da tabela de Data Mining

A tabela de Data Mining (DM) -data_mining_table_stock- contém o essencial de toda a nossa base de dados mas desnormalizada. (é o que os algoritmos de DM precisam para manipular os dados). Criar esta tabela não foi uma tarefa tão simples quanto isso. Do ponto de vista do negócio (ie: evolução dos mercados financeiros) é preciso ser criterioso para saber que dados fazem sentido aqui pôr, além disso, é preciso pensar em qual será a melhor forma de colocar os dados de modo a “facilitar a vida” ao algoritmo de DM. Por exemplo, optámos por calcular sempre valores relativos e não valores absolutos e sempre que possível usámos rácios.

Do ponto de vista técnico a criação desta tabela também não foi simples (tem mais de 150 campos). Uma parte da criação da tabela até é feita com uma pequena rotina que fizemos para construir nomes de campos. O preenchimento da tabela em si é um gigantesco join entre as tabelas que temos. (mas nem sempre é possível fazer join, depois há uns updates).

Criámos 2 stored procedures para criar esta tabela. Um adiciona um índice completo e outro um título específico entre uma data inicial e uma data final.

Escolha do algoritmo de classificação

O classificador usado foi essencialmente o C5.0 (árvores de decisão). No início foram feitos testes comparativos com redes neuronais, algoritmos de regressão linear e algoritmos de árvores de decisão, revelando estes últimos resultados mais adequados para a modelação do problema.

Classificador C5.0

Um classificador C5.0 funciona por fraccionamento do conjunto dados de treino pelos atributos que em cada conjunto de dados permitem o máximo ganho de informação. Cada sub conjunto definido pelo fraccionamento anterior é em seguida fraccionado recorrendo aos mesmos critérios, num processo recursivo, até que não seja possível mais fraccionamento. Por último os últimos ramos da árvore são analisados, e aqueles que não contribuem significativamente para o ganho de informação são removidos. A decisão quando à variável objectivo surge nas folhas da árvore de decisão, com o

respectivo nível de confiança calculado com base no número de casos que verificaram o valor escolhido para a variável objectivo nesse ramo terminal da árvore.

O algoritmo C5.0 pode manipular atributos discretos bem como números reais, permitindo grande flexibilidade e adaptabilidade a variados problemas de Data Mining. Contudo, as variáveis objectivo devem ser de natureza discreta.

Treino de Classificadores

A fase de treino dos modelos foi essencial para determinar as parametrizações do classificador C5.0 de forma a permitir os melhores resultados relativamente aos dados de validação. Os testes efectuados envolveram a iteração de diferentes configurações nos seguintes conjuntos de variáveis:

- Atributos de entrada
- Discretizações das variáveis objectivo
- Variáveis objectivo
- Tempo de treino
- Parâmetrização do classificador C5.0
- Tempo de validação

Atributos de entrada

Quanto às variáveis de entrada para treino dos modelos, foi possível chegar às seguintes metodologias de triagem: Verificou-se que as variáveis que representam valores absolutos e com estrita relação com a uma dada situação no mercado bolsista levam frequentemente ao problema do *overfitting* dos modelos para o período de treino e á consequente falta de eficácia em períodos futuros. Pelos mesmos motivos foram retiradas variáveis como o dia do ano e a semana do ano, já que, com a presença destas, mais facilmente o modelo estaria viciado para coincidências casuais no período de treino.

Discretizações das variáveis objectivo

As diferentes discretizações das variáveis objectivo foram testadas, e verificadas quais as mais fiáveis considerando os resultados do modelo para o período de validação.

Variáveis objectivo

Os vários modelos para as diferentes variáveis objectivo foram comparados revelando resultados distintos.

Tempo de treino

O tempo de treino é essencial para a captura dos padrões mais significativos que levam a uma determinada evolução das cotações. Foram testados vários períodos tendo em atenção que, períodos demasiado longos levam ao treino de modelos com comportamentos eventualmente desactualizados

e não aplicáveis para o tempo de validação. Por outro lado, tempos de treino demasiadamente curtos levam a uma maior dificuldade na extracção das regras de comportamento visto que um período curto pode ser pobre no conteúdo de informação.

Parâmetrização do classificador C5.0

A parametrização do Classificador C5.0 foi dirigida no sentido de produzir modelos genéricos extraíndo as regras mais significativas para cada título, ou seja, as regras com um número de casos considerado suficiente para que a probabilidade de se tratarem de meros acasos seja minimizada.

Tempo de validação

O tempo de validação, naturalmente, foi sempre posterior ao período de treino, sendo que os modelos foram testados para o ano seguinte ao período e treino.

Inicialmente procurou-se gerar modelos para previsão dos objectivos propostos para um conjunto de títulos pertencentes a um dado índice. Testou-se a criação de modelos para o índice americano NASDAQ constituído por 100 empresas e para o qual estão disponíveis volumes consideráveis de dados.

Os resultados obtidos revelaram percentagens de classificações correctas muito aceitáveis para o período de treino do modelo. Já relativamente a um período futuro os resultados foram bastante fracos, a rondar o que se obteria com a escolha aleatória das subidas e descidas. Deste modo, foi escolhida uma abordagem de treino de modelos individuais para cada título na tentativa de aproveitar melhor padrões comportamentais intrínsecos a cada título. Para o treino de modelos para cada título optou-se por considerar para esta fase apenas aqueles pertencentes ao índice NASDAQ.

O treino de modelos para cada título reduziu dramaticamente o número de casos de treino para cada modelo. Sendo assim, os intervalos de tempo usados para treino dos modelos tiveram que ser superiores aos usados para o modelo global de forma a permitir a captura da informação relevante.

Treino e análise dos modelos de Data Mining para um caso de exemplo

O tipo de análise efectuada aos modelos é em seguida demonstrado com base nas ferramentas do *Clementine*®.

Considerando a título de exemplo o modelo gerado para as acções da Microsoft seguimos o seguinte processo:

Para gerar um modelo usando a ferramenta *Clementine*®, em primeiro lugar definem-se as etapas de tratamento de dados. Para o presente trabalho, a fonte de dados foi a base de dados *SQL Server*®.

Em seguida procede-se à discretização da variável objectivo, neste caso a evolução da cotação ao fim de 7 dias, (*STK_EVOL_VAL_7*) para duas classes, respectivamente, evolução positiva e

negativa. Depois foram definidos os atributos utilizados para treino do classificador usando o nó *Type* do *Clementine*®.

Depois de gerado o modelo, é possível inserir o nó modelo na *stream*, e analisar os resultados para o período que se desejar de acordo com a query SQL definida no nó de obtenção de dados (Figura 13).

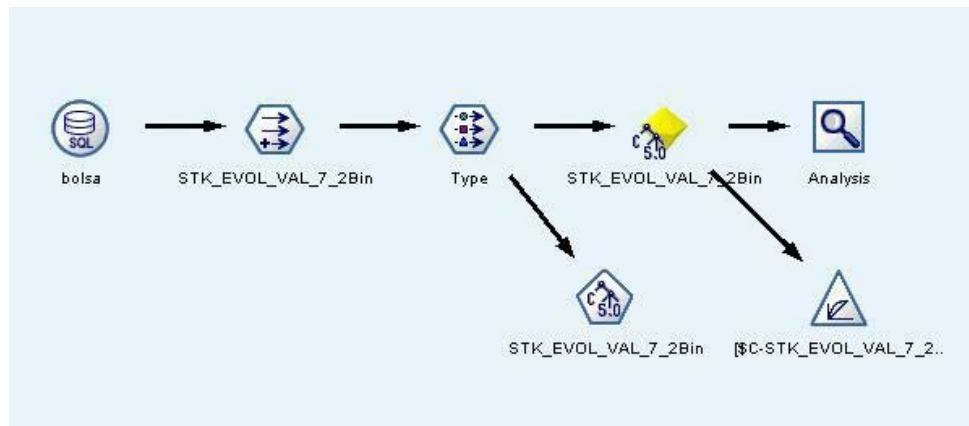


Figura 14: Aspecto da *stream* para treino e teste do modelo para a Microsoft no *Clementine*®

Para o treino de modelos individuais para cada título considerou-se, depois de testes exaustivos, um período de dados de 5 anos. Enquanto que, para o período de validação foi usado o período de 1 ano seguinte ao período de treino.

Um dos parâmetros para o classificador C5.0 é o número mínimo de casos para a criação de um ramo novo da árvore de decisão. Depois de vários testes verificou-se que números elevados levam à captura das regras mais genéricas e em consequência, uma melhor adaptação do modelo a períodos diferentes dos períodos de treino. Sendo assim foram escolhidos como valor mínimo 45 casos por ramo da árvore.

Assim, procedeu-se ao treino do modelo para o período 1998 – 2002 tendo-se obtido a seguinte árvore de decisão (Figura 14):

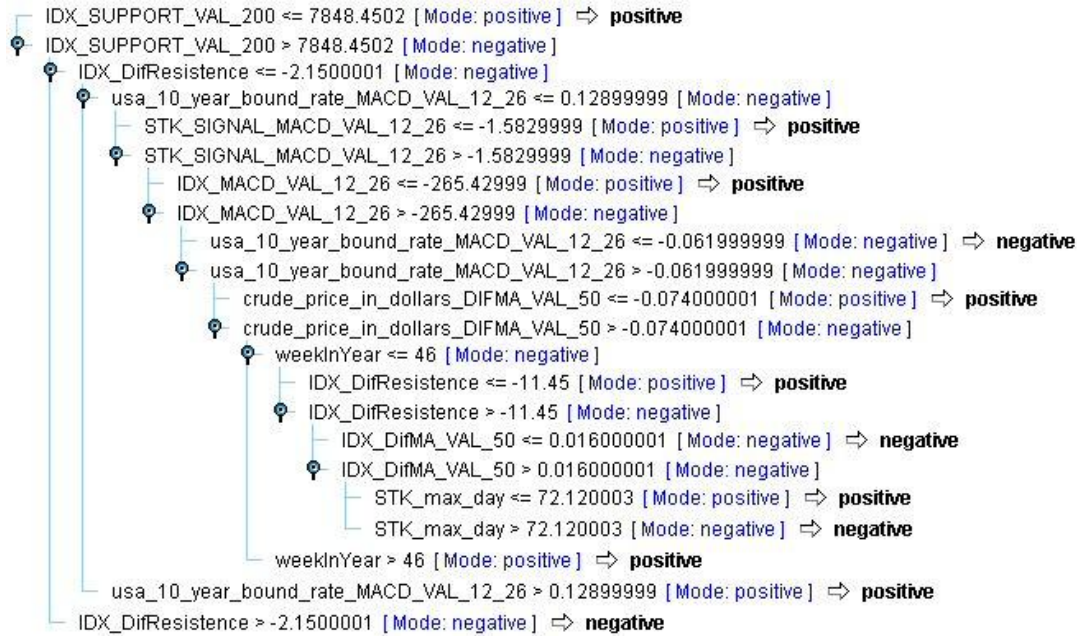


Figura 15: Árvore de decisão para o modelo da Microsoft treinado para o período de 1998 a 2002

Como se pode concluir pela análise da árvore de decisão, verifica-se que foram usadas variáveis relativas à evolução do índice NASDAQ, nomeadamente, o valor de suporte calculado para o intervalo de 200 dias. Foram igualmente utilizados os indicadores para as cotações da Microsoft e os indicadores para as variáveis de economia mundiais tal como o preço do petróleo.

Analisando a qualidade do modelo para os casos de treino, verificamos que a precisão é na ordem dos 75% de casos correctos para o conjunto de previsões de evolução negativa e positiva.

Mas fazendo a análise para o período de validação (ano de 2003), ou seja, um período de dados que não foi utilizado no treino do modelo, verifica-se que a precisão foi muito abaixo da obtida para o período de treino mas permanecendo com um ganho de quase 7% face ao palpite aleatório (50%).

Comparação dos valores reais com os valores estimados

Correct	143	56,75%
Wrong	109	43,25%
Total	252	

Matriz de coincidências (as linhas mostram os valores reais)

	negative	Positive
negative	77	42
positive	67	66

Avaliação de performance

negative	0.124
positive	0.147

Relatório de valores de confiança

Range	0.635 - 0.877
Mean Correct	0.787
Mean Incorrect	0.796
Always Correct Above	0.877 (0% of cases)
Always Incorrect Below	0.635 (0% of cases)
90% Accuracy Above	Never reached requested level
2.000 Fold Correct Above	Never reached requested level

Quanto à curva de ganho (Figura 15), pode observar-se a percentagem de previsões do modelo relativamente a evoluções positivas, necessários para capturar x% dos casos reais. A linha preta representa a precisão do palpite aleatório.

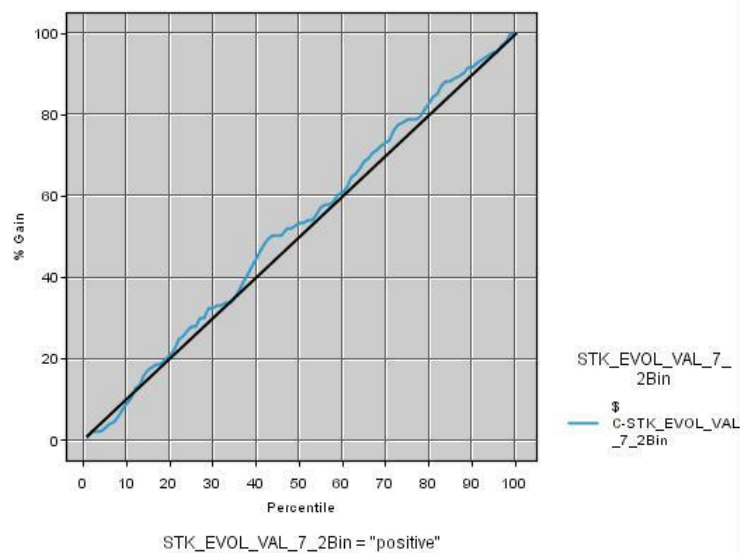


Figura 16: Gráfico de ganho relativo ao modelo para a Microsoft no período de validação de 2003

Automatização da geração de modelos no Clementine®

Depois de termos ganho experiência a trabalhar com o Clementine em modo interactivo fizemos uma pequena simulação mais alargada em Excel (ver anexo Simuladores em Excel).

Estávamos com o problema de como exportar os resultados de modo a poderem ser usados automaticamente dentro da nossa aplicação. Concluímos que a melhor maneira seria por os resultados como uma tabela da nossa base de dados.

Ainda nos restava o problema de alterar dinamicamente uma stream de execução do Clementine. Para isso teríamos de as manipular em runtime. Numa primeira fase descobrimos que os ficheiros str do Clementine® são na realidade um ficheiro xml com compressão Zip. No site do Clementine® até

encontrámos um reader para estes ficheiros xml. Ainda pensámos em alterar a stream directamente em xml e depois voltar a comprimi-la, mas rapidamente nos apercebemos que isso não seria nada fácil.

Assim virámo-nos para a linguagem de *scripting* do *Clementine*®. Depois de meio-dia a estudá-la (infelizmente o manual não é muito explícito neste aspecto nem há muitos exemplos), fizemos um *script* (dm_model.clm) que, manipulando um stream mais virado para ser processado automaticamente (dm_model.str) nos permite gerar, sem qualquer intervenção “humana” os resultados dos modelos de Data Mining.

O *script* leva 3 parâmetros: a empresa para a qual queremos gerar o modelo, o período de treino, e o período de teste (no futuro relativamente ao período de treino). O script gera os 6 modelos de evolução para a empresa em causa. (a previsão de evolução a 1 dia, 3 dias, 7 dias, 15 dias, 30 dias e 60 dias).

Depois de gerados os 6 modelos (criados no período de treino), estes são corridos com dados novos (do período de teste). O resultado de correr estes modelos é posto na tabela **data_mining_stock_results** (ver anexo dicionário de dados) que é inserida na nossa base de dados.

Finalmente conseguimos integrar a chamada do script dentro do C# como uma chamada ao sistema.

```
Process.Start(clementineExeFileName, parameters).WaitForExit();
```

Avaliação dos modelos gerados pelo Clementine®

Os testes efectuados para aferir a capacidade predictiva dos modelos foram os seguintes:

Teste para o período de treino

Este teste, apesar de ser importante pode apresentar resultados enganadores que revelam a presença de *overfitting* para o conjunto de treino. Ou seja, quando um modelo se adequa muito bem ao período de treino podemos estar na presença de aprendizagem com regras que apenas existem no período de treino e que no tempo futuro não possuímos qualquer garantia ou fundamento para que aconteçam de novo.

Teste para um período futuro ao período de treino

Este é o teste mais importante na aferição da capacidade predictiva dos modelos. O conjunto de dados não pertence ao período de treino do modelo, logo aqui podemos verificar se as regras descobertas pelo classificador são realmente válidas e generalistas o suficiente para não estarem vinculadas apenas ao período de treino do modelo.

Análise Global para o conjunto de modelos do NASDAQ

Os modelos gerados para o conjunto de títulos do NASDAQ revelou-se pouco preciso mesmo para o conjunto de treino. Para isto pensamos que contribuiu o facto dos comportamentos para cada título serem bastante distintos entre si sendo difícil encontrar uma receita que se adapte a todos eles com os dados de que dispomos.

Confiança nos níveis de confiança

Uma das formas de procurar extrair a informação dos modelos com maior confiança pode ser considerar apenas as previsões às quais são atribuídos níveis de confiança acima de determinados valores. Assim torna-se importante verificar se de facto a precisão aumenta com o aumento dos níveis de confiança mínimos.

As figuras seguintes mostram os resultados desta análise para a totalidade dos modelos gerados para o índice NASDAQ.

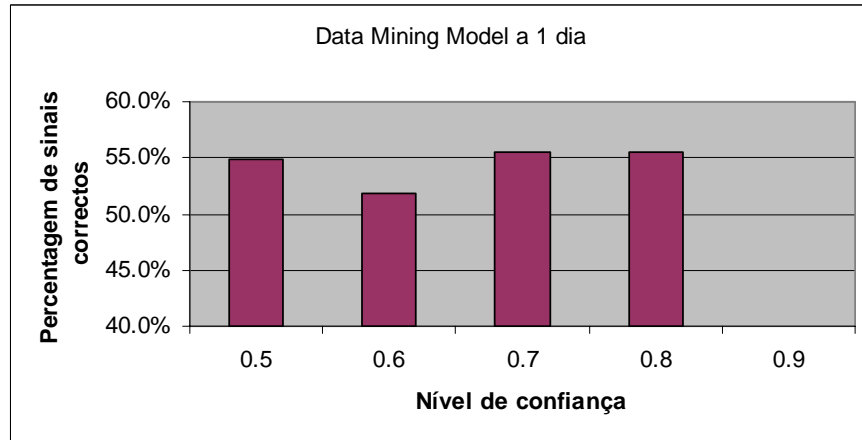


Figura 17: Evolução da precisão do modelo de evolução a 1 dia face a diferentes valores mínimos de confiança

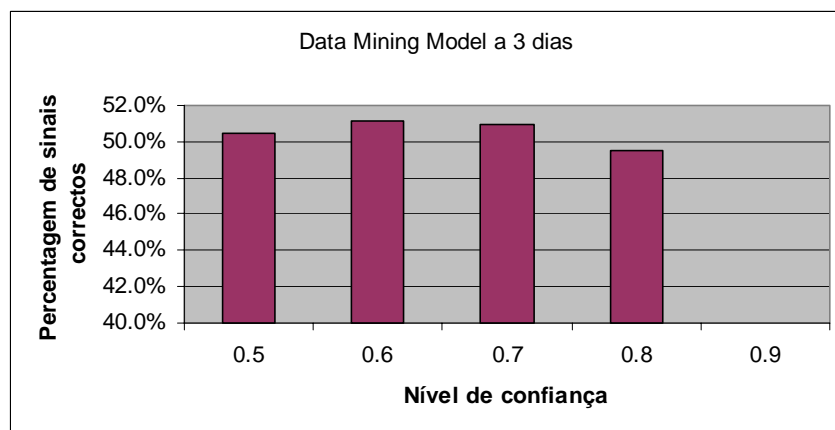


Figura 18: Evolução da precisão do modelo de evolução a 3 dias face a diferentes valores mínimos de confiança

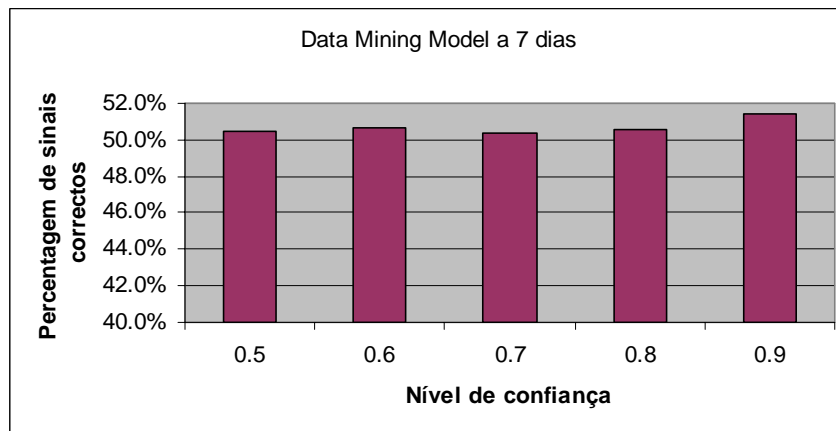


Figura 19: Evolução da precisão do modelo de evolução a 7 dias face a diferentes valores mínimos de confiança

Pelos gráficos apresentados, pode concluir-se que a precisão dos modelos obtidos quando analisada globalmente não apresenta ganhos significativos face ao palpite aleatório. Os ganhos de precisão, nos melhores casos, mal chegam a ultrapassar os 5% face aos 50% obtidos aleatoriamente.

Também se pode concluir que, globalmente, o aumento dos níveis de confiança não melhora visivelmente a precisão dos modelos.

No entanto, seria de esperar que o comportamento de alguns títulos seja mais passível de ser interpretado por Data Mining do que outros. Logo, analisando globalmente, essas melhores prestações ficam escondidas face à média global. Seguindo este raciocínio procurámos analisar o conjunto dos 10 melhores modelos considerando a sua adaptação a períodos posteriores ao período de treino.

10 Melhores modelos

Os 10 melhores modelos foram seleccionados utilizando a componente da aplicação de suporte à decisão na bolsa, *ModelAccuracy* e considerando, para cada um, os resultados de dois modelos treinados para os períodos de 1997 a 2001 e 1998 a 2002 e para períodos de validação correspondentes aos anos de 2002 e 2003 respectivamente.

Os resultados desta análise podem ser observados no anexo 5.

Considerando os melhores modelos, observam-se precisões bem acima dos 50% frequentemente com valores acima de 70%. Em geral obtiveram-se melhores resultados para os modelos para períodos mais longos.

Não podemos contudo deixar de fazer a ressalva de que, os melhores modelos foram escolhidos tendo em conta os resultados para o período posterior ao de treino, logo, considerando uma aplicação para suporte à decisão de investimento na bolsa em condições, reais naturalmente não teríamos acesso aos valores futuros e não seria possível determinar desta forma quais os melhores modelos.

Pode-se contudo analisar, se os títulos com melhores modelos tende a continuar assim em períodos de validação consecutivos e os correspondentes períodos de treino.

Análise Global da Aplicação de Data Mining

Quando à componente de Data Mining, os modelos desenvolvidos para alguns casos de títulos do índice NASDAQ revelam resultados muito satisfatórios. Um raciocínio natural seria considerar que a evolução destes títulos seria mais facilmente entendida pelos respectivos modelos. Mas essa assumpção pode não ser assim tão clara. Nada nos pode garantir que um modelo sobre um mesmo título quando usado para prever de facto a evolução futura, venha a obter resultados igualmente satisfatórios.

A aplicação de Data Mining á bolsa é de extrema complexidade dada a variedade de metodologias que podem ser elaboradas e a complexidade do problema em si. Dificilmente numa primeira tentativa se conseguem resultados óptimos. No entanto, a plataforma aplicacional desenvolvida neste projecto permite facilmente alteração dos atributos usados para treino dos modelos bem como a utilização de outros classificadores, logo, mesmo que os resultados não sejam os ideias, muito trabalho pode ainda ser efectuado sobre esta plataforma para atingir eventualmente resultados ainda mais interessantes.

Descrição da Aplicação de suporte à decisão

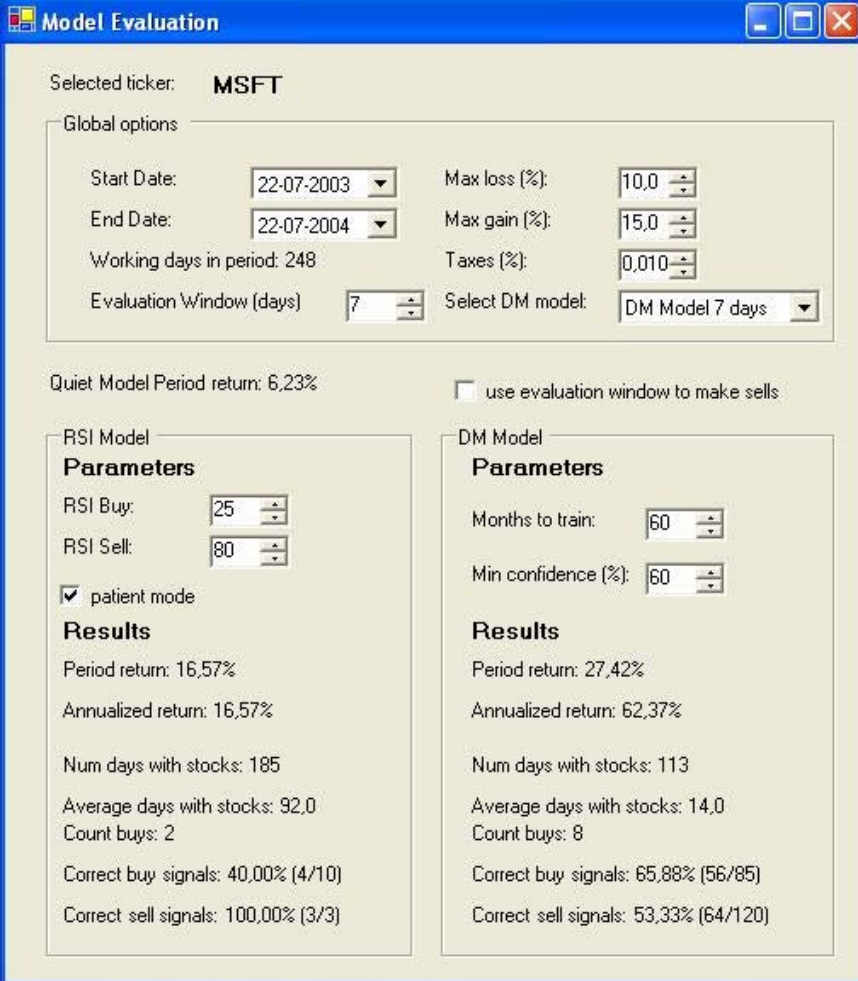
Módulo de Simulação

Para avaliar a qualidade do nosso modelo fizemos um simulador bastante flexível.

Escolhemos qual o período em que queremos testar e depois várias opções.

Antes de comparar modelos elaborados é importante saber com o que devemos comparar.

Devemos comparar com a evolução da bolsa no período referido. (*Quiet Model Period Return*)



Selected ticker: **MSFT**

Global options

Start Date: 22-07-2003 Max loss (%): 10,0

End Date: 22-07-2004 Max gain (%): 15,0

Working days in period: 248 Taxes (%): 0,010

Evaluation Window (days): 7 Select DM model: DM Model 7 days

Quiet Model Period return: 6,23% use evaluation window to make sells

RSI Model Parameters

RSI Buy: 25

RSI Sell: 80

patient mode

Results

Period return: 16,57%

Annualized return: 16,57%

Num days with stocks: 185

Average days with stocks: 92,0

Count buys: 2

Correct buy signals: 40,00% (4/10)

Correct sell signals: 100,00% (3/3)

DM Model Parameters

Months to train: 60

Min confidence (%): 60

Results

Period return: 27,42%

Annualized return: 62,37%

Num days with stocks: 113

Average days with stocks: 14,0

Count buys: 8

Correct buy signals: 65,88% (56/85)

Correct sell signals: 53,33% (64/120)

Figura 20: Interface gráfica para teste de simulações

O que o nosso simulador genérico faz é, dada uma sequência de sinais de compra e venda executa-os e mostra o resultado e algumas estatísticas.

O modelo mais simples é o RSI em que os sinais de compra e de venda são gerados quando o indicador RSI relativo à acção seleccionada chega aos valores escolhidos. O RSI de compra vale tipicamente entre 15 e 35 e o de venda entre 70 e 90.

O RSI em modo paciente espera que a acção atinja os valores escolhidos mas em subida (na compra) e em queda (na venda). Ou seja, em modo paciente só compramos no momento em que o RSI já recuperou acima de X (e no período anterior esteve abaixo de X). O modo “impaciente” compra mal desça abaixo de X. O mesmo se passa analogamente para a venda.

Além do RSI (que não precisamos do DataMining para o gerar) temos o modelo de Data Mining. Aliás, podemos escolher qual dos 6 modelos de DataMining queremos. No modelo de Data Mining há 2 parâmetros importantes: Qual a confiança mínima com que fazemos as compras e vendas, e qual a duração do período de treino do modelo. Se exigirmos mais confiança vamos fazer menos compras/vendas mas supostamente estas seriam mais certas (isto nem sempre é verdade... pelas experiências que fizemos nem sempre maior confiança implica maior certeza na previsão). O outro parâmetro é o tempo de treino em meses. Estes meses são imediatamente antes da data seleccionada.

Temos ainda dois parâmetros que ajudam a controlar. A perda máxima a partir da qual, mesmo que o modelo não esteja a mandar vender, vendemos para proteger o capital (pode ser boa estratégia ser tolerante às perdas...mas ser demasiado tolerante também pode ser muito mau...) e o lucro que, assim que atingido, nos damos por satisfeitos e não esperamos pelo sinal de venda. (que podia vir mais tarde com um valor mais alto...mas também mais baixo!)

Também consideramos as taxas. A taxa é a comissão cobrada pela corretora por cada compra/venda.

Os modelos de data mining a mais de um dia podem parecer um bocado idiotas (ainda que até se possam comportar bem!): compramos hoje se o modelo diz que daqui a X dias sobe, ou vendemos hoje se o modelo diz que daqui a X dias desce.

A evaluation window é a janela de tempo para a qual é avaliado se o modelo acertou ou não (correct buys e correct sells). O que faz sentido é que o número de dias na evaluation window seja o mesmo número de dias que o modelo de DM.

Quando se liga a opção “use evaluation window to make sells” são ignorados os sinais de venda gerados pelo modelo. A venda passa a ser efectuada “evaluation days” depois da compra.

Módulo de análise técnica



Figura 21: gráfico com o histórico do último ano da Microsoft visto com alguns indicadores

Nesta forma temos bastante informação de apoio ao investidor. Quando escolhemos um ticker é mostrado o gráfico das cotações para o período seleccionado (por default um ano) em formato candlestick. Pode-se também escolher que indicadores se deseja ver. Alguns dos indicadores têm uma escala diferente da série da cotação e portanto aparecem num novo eixo (como o RSI que dá sempre um valor entre 0 e 100). Outros têm valores na mesma escala da série de cotação (por exemplo, as médias móveis). Quando se liga a opção “serie details” é mostrada informação relativa ao dia em que o cursor do rato está. Além dos indicadores é mostrado quando houve dividendos e de que valor foram, quando houve splits e quais os rácios e quando foram detectados padrões e quais os padrões detectados.

Há ainda o botão “download latest data” que permite ir à web buscar os últimos dados relativos à série de cotações. (a informação relativa aos indicadores é também actualizada mas pela aplicação).

É ainda possível fazer zoom dentro do gráfico desenhando um rectângulo da esquerda para a direita com o mouse dentro do gráfico. Para fazer unzoom faz-se um rectângulo da direita para a esquerda.

Diferenças entre visualizar com candlesticks e com fecho ajustado



Figura 22: gráfico com o histórico de sempre da Microsoft visto com candlesticks

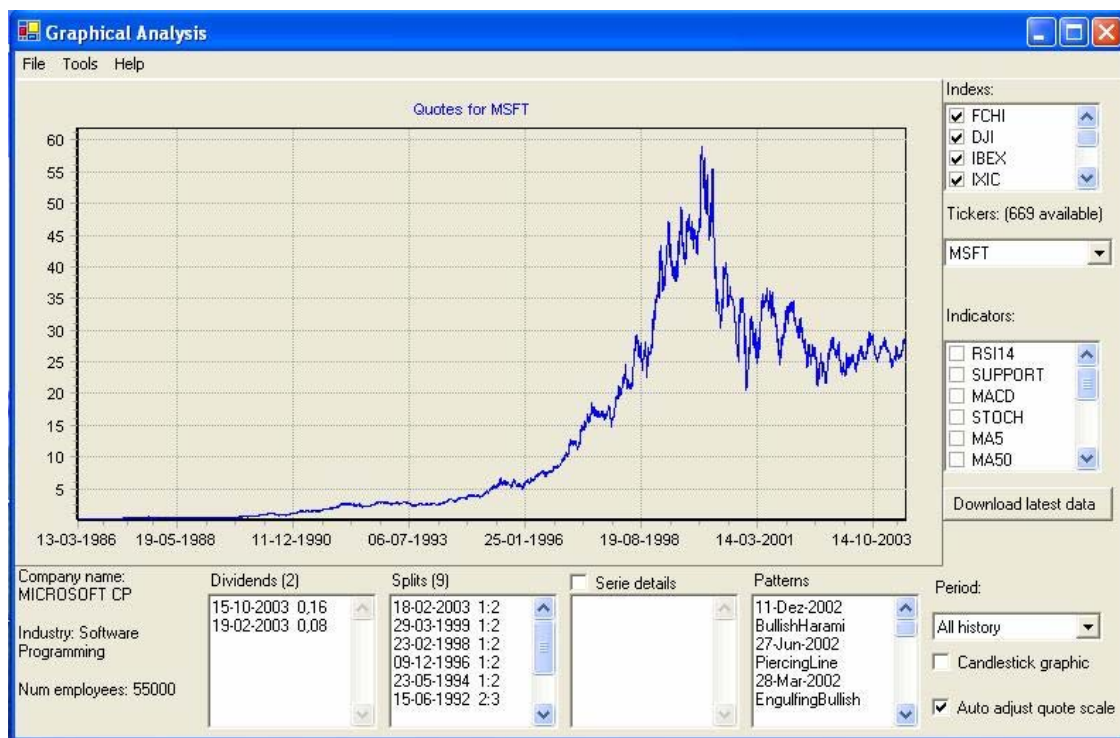


Figura 23: gráfico com o histórico de sempre da Microsoft visto com adjusted closing

Na visualização com candlesticks são mostrados 4 valores: Open, High, Low, Close. Na visualização com adjusted closing é apenas mostrado este campo. Para a data de hoje o close e o adjusted closing são sempre iguais. A diferença entre ambos é que o adjusted closing tem em conta os eventuais splits/mergers que tenham ocorrido ao longo do tempo.

Por exemplo: em 18 de Fevereiro de 2003 a Microsoft fez um split de 1 para 2. Cada acção de 18 de Fevereiro de 2003 passa a valer 2 em 19 de Fevereiro de 2003. Para isto acontecer o valor da acção cai, em 19 de Fevereiro, para metade, naturalmente. No campo adjusted closing não há descontinuidade. O adjusted closing é sempre recalculado para todo o passado sempre que há um fusão/divisão ou distribuição de dividendos.

Ao analisar o histórico o que faz sentido é ver com adjusted closing para ter uma ideia real do valor da acção em cada tempo. A visualização com candlesticks dá mais informação mas pode induzir em erro. Ao ver uma queda abrupta podemos achar que foi por causa de um split mas pode ter sido uma queda natural. De referir que, embora o rácio mais frequente seja 1:2, por vezes há rácios de split bem estranhos que poderiam ser a causa de descidas no gráfico.

Módulo de análise da precisão das previsões dos modelos de Data Mining

Rank	Ticker	Accuracy (%)	Count
1)	GRMN	87,76%	(86/98)
2)	XRAY	76,79%	(182/237)
3)	FHCC	76,34%	(100/131)
4)	NXTL	76,05%	(289/380)
5)	AMZN	75,24%	(319/424)
6)	CTAS	73,49%	(352/479)
7)	EBAY	72,61%	(342/471)
8)	AMGN	71,10%	(251/353)
9)	JNPR	69,50%	(303/436)
10)	CHIR	68,00%	(153/225)
11)	PCAR	66,08%	(261/395)
12)	PSFT	63,83%	(120/188)
13)	ROST	62,92%	(151/240)
14)	MEDI	62,82%	(196/312)
15)	YHOO	61,58%	(226/367)
16)	MCHP	60,78%	(234/385)
17)	BBBY	60,65%	(225/371)
18)	FLEX	60,14%	(175/291)
19)	INTU	60,06%	(212/353)
20)	BMET	60,00%	(273/455)

Figura 24: Form de avaliação da precisão do modelo de DM

Depois dos modelos de DM criarem as tabelas de resultado (que têm a previsão e a confiança na previsão) podemos fazer uma análise para determinar quais são os títulos em que o modelo acerta melhor.

Nesta form escolhemos qual o modelo de Data Mining a considerar e qual a confiança mínima com que queremos gerar os sinais de compra/venda.

O resultado é a percentagem de sinais gerados correctamente e o número de casos com que foi obtida essa percentagem.

Módulo de análise técnica

Figura 25: Form de pesquisa de tickers por condição

Nesta form podemos pesquisar quais os tickers que verificam a união de várias condições.

Primeiro escolhemos qual a data a considerar (X) e depois quantos dias úteis antes dessa data (N). As pesquisas que se farão estão confinadas ao intervalo [X-N, X]. Também temos de escolher quais os índices em que queremos pesquisar.

As condições que se pesquisam são: padrões de candlestick, valor do RSI e previsões de qualquer dos modelos de Data Mining.

Mais concretamente vamos listar todos os tickers em que foi detectado um padrão no intervalo referido, ou que têm o RSI abaixo ou acima de um valor escolhido ou que o modelo de data mining seleccionado faz uma previsão (de subida ou descida) com confiança maior ou igual à escolhida.

Conclusão

Apreciação Crítica do Trabalho Desenvolvido

O trabalho foi desenvolvido num intervalo de tempo muito limitado (apenas 2 meses), no entanto, apesar dessa limitação foi possível elaborar uma aplicação com ferramentas de análise técnica que sem dúvida consistem numa mais valia no suporte à decisão em investimento na bolsa de valores.

A integração dos principais indicadores de análise técnica sobrepostos às evoluções diárias das cotações para um vasto conjunto de títulos das principais bolsas internacionais fornece um completo ambiente de análise.

A recolha em tempo real de novos dados permite que a base de dados seja facilmente actualizada com os novos valores diários.

A componente de aplicação de Data Mining à bolsa de valores permitiu numa primeira abordagem alcançar resultados, em alguns casos bastante interessantes, com precisões das previsões de evolução na bolsa bem acima das obtidas com palpites aleatórios.

Nesta primeira abordagem à problemática do Data Mining sobre a bolsa de valores foi possível adquirir conhecimentos sobre os seus comportamentos e sobre várias técnicas de análise. Com a experiência adquirida, neste momento seria possível elaborar alterações à metodologia adoptada, tais como o acrescento de indicadores alternativos ou alteração dos formatos para os já desenvolvidos. Nesta perspectiva, a plataforma aplicacional desenvolvida permite facilmente o treino e integração de novos modelos de Data Mining bem como a utilização de diferentes atributos de treino.

Para que fosse possível um planeamento mais ponderado e uma análise mais profunda sobre a componente técnica da bolsa seria proveitoso a elaboração do presente projecto com menos restrições temporais. No entanto, por motivos profissionais, académicos e pelos imprevistos iniciais sobre o alvo de aplicação do estágio não nos foi possível dispor de todo o tempo que gostaríamos. Apesar de tudo, e como é hábito em estágios em Engenharia Informática, tencionamos continuar o trabalho desenvolvido já que muitos caminhos existem ainda por explorar e a plataforma desenvolvida a isso convida.

Conhecimentos adquiridos

Durante o estágio foi possível estabelecer um primeiro contacto com a actividade de uma empresa de consultoria informática como é a Novabase, e em particular, com a actividade desenvolvida na área de Business Intelligence.

Com a elaboração deste projecto, adquirimos fortes conhecimentos práticos na aplicação de Data Mining nomeadamente na utilização das ferramentas *Clementine*® e Weka. Inclusivamente o uso da linguagem de scripting (CLEM) do *Clementine*®

Consolidaram-se conhecimentos nas áreas de Bases de Dados, nomeadamente na utilização de bases de dados SQL Server e programação de procedures em transact SQL.

Neste estágio foi possível ter o primeiro contacto com a linguagem C# usada para o desenvolvimento integral da presente aplicação.

Trabalho Futuro

No seguimento do trabalho desenvolvido foi possível observar que seria possível implementar algumas melhorias que poderiam eventualmente produzir melhores resultados ao nível dos modelos, nomeadamente a utilização de novos indicadores na componente de análise técnica e na componente de Data Mining.

Aplicabilidade dos conhecimentos adquiridos na LEI

Este projecto versa várias áreas, assim tivemos oportunidade de pôr em prática o que aprendemos em diversas cadeiras da licenciatura.

Nomeadamente: Programação I e II (em geral), Redes (Recolha de dados), Linguagens Formais e Autómatos, Linguagens e Técnicas de Programação Avançadas (Expressões Regulares para tratamento dos dados recolhidos), Bases de Dados I (Transact SQL, SQL Server), Análise de Sistemas (Desenho geral do sistema), Probabilidades e Estatística (Interpretação dos resultados de Data Mining), Algoritmos e Estruturas de Dados, Introdução à Inteligência Artificial (Reconhecimento de Candlesticks e alguma compreensão do Data Mining), Linguagens de Programação I (Estrutura de classes no módulo de simulação de modelos).

Calculamos que as cadeiras de Base de Dados e Data Warehousing e Aprendizagem Automática seriam muito úteis neste projecto mas foram cadeiras opcionais que nenhum de nós escolheu fazer. A cadeira de Data Mining também deveria ser muito útil mas só é possível fazê-la no mestrado.

Apreciação do Estágio

O nosso estágio teve 2 fases distintas. De 5 de Abril a 24 de Maio e de 24 de Maio a 23 de Julho.

Inicialmente o objectivo do estágio era diferente. O título era: “**Detecção de fraude na segurança social usando tecnologias de Data Mining e Data Warehousing**”. Era altamente motivante estar a trabalhar nesta área pois estaríamos a ajudar a resolver um problema muito relevante para a sociedade.

Durante os primeiros 2 meses foi nesse projecto que estivemos a trabalhar. O nosso orientador na NBBI, eng^o Pedro Moura, explicou-nos que muito possivelmente iríamos fazer uma prova de valor

no Instituto de Informática e Estatística da Segurança Social (IIES). Para nos prepararmos para isso durante as primeiras semanas estivemos a familiarizarmos com o SQL Server 2000 (com o qual nunca tínhamos trabalhado) e também com o Weka, um software de Data Mining open source. Fizemos algumas experiências com bases de dados que nos facultaram para percebermos o processo de Data Mining e a ferramenta. Treinámos com problemas diversos como: Previsão de fraude nas declarações do IVA, Previsão de saída de clientes de um cartão de crédito, Previsão de saída de clientes de uma operadora de telecomunicações. Com cada um destes problemas estivemos uma ou duas semanas.

Dia 24 de Maio de 2004 soubemos que não iríamos para o IIES. Este preferiu implementar o projecto internamente apenas com a supervisão de um consultor da Novabase. Assim tivemos de arranjar um projecto novo.

Após uma conversa com o nosso orientador na FCT e depois com o orientador da Novabase surgiu a ideia que deu origem ao projecto que agora apresentamos: “**Sistema de Apoio à Decisão para investimento na Bolsa de Valores usando Data Mining**”. A partir dessa primeira reunião tivemos muito trabalho intensivo pela frente e aproveitámos boa parte da experiência adquirida.

Teríamos naturalmente preferido que não tivesse havido este imprevisto mas a vida profissional é mesmo assim.

Bibliografia

- [1] Usama M. Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, and Ramasamy Uthurusamy, *Advances in Knowledge Discovery and Data Mining* – AAAI Press, 1996.
- [2] Cabena, Hadjinian, Staedler, Verhees, Zanasi, *Discovering Data Mining from Concept do Implementation*, Prentice Hall, 1997.
- [3] Ian H. Witten, Eibe Frank, *Data Mining, Practical Machine Learning Tools and Techniques with Java Implementation*, San Francisco, California, Morgan Kaufmann Publisher,
- [4] Nison, Steve, *Japanese Candlestick Charting Techniques"* , 1991
- [5] web site <http://finance.yahoo.com>
- [6] web site <http://www.incrediblecharts.com>
- [7] web site <http://www.caldeiraodebolsa.com>
- [8] web site <http://www.inverline.com>
- [9] web site <http://www.investorhome.com/mining.htm>

Anexo 1 - Descrição dos ficheiros de inicialização da base de dados

Descrição do ficheiro de sectors.txt:

Primeira linha: numero de sectores

Para cada sector a primeira linha está em branco, e a segunda contém o nome do sector. A terceira linha contém o número de industrias, M. As seguintes M linhas têm o nome de cada industria pertencente ao referido sector.

Descrição do ficheiro indexs.txt:

Primeira linha: número de indexs

Para cada índice temos:

Primeira linha: em branco

Segunda linha: nome do índice:

Terceira linha: código do índice

Quarta linha: país do index

Quinta linha: quantos títulos há neste índice.

Para cada título temos 2 linhas. A primeira é o ticker e a segunda é o nome completo.

Descrição do ficheiro world_indicator_codes.txt:

Primeira linha: número de indicadores mundiais conhecidos, N.

Linha 2 a N: Cada linha tem o nome do indicador.

Descrição do ficheiro currencies.txt:

Primeira linha: número de moedas conhecidas, N.

Linha 2 a N: Cada linha tem o nome da moeda.

Descrição do ficheiro countries.txt:

Primeira linha: número de países

Para cada país temos:

Primeira linha em branco

Segunda linha: nome do país

Terceira linha: nome da moeda do país

Descrição do ficheiro candlepatterns.txt:

Primeira linha: número de padrões conhecidos, N

Linha 2 a N: Cada linha tem o nome do padrão.

Descrição do ficheiro brentbarril.txt:

Cada linha do ficheiro tem 2 campos: data e valor do barril de brent (em USA dolares) nessa data. A separação é feita por ponto e virgula. A data está no formato dd-mm-yyyy. O valor do barril é um número real mas nalguns casos pode ser a string 'NULL'

Descrição das fontes de dados

De seguida listamos as fontes de dados. De notar que para tirar exactamente a informação que precisávamos tivemos de fazer parsing ao conteúdo da página procurando expressões regulares.

Informação relativa a uma empresa:

Encontramos a informação relativa a uma empresa no seguinte url:

"http://finance.yahoo.com/q/pr?s="+ticker;

Em que ticker é o código da companhia.

Deste url apenas aproveitamos 2 valores: a indústria em que a empresa se insere e o número de empregados.

Informação relativa ao indicador USA 10 year bound rate:

Encontramos a informação relativa a este indicador no seguinte url:

"http://ichart.yahoo.com/table.csv?s=^TNX&ignore=.csv";

A este url é preciso concatenar a data a partir da qual queremos os dados.

Se quisermos a série toda fazemos por exemplo:

"http://ichart.yahoo.com/table.csv?s=^TNX&ignore=.csv"&a=00&b=1&c=1930";
//a for month-1, b for day, c for year

Se quiséssemos apenas os dados desde 15 de Maio de 2003, faríamos:

"http://ichart.yahoo.com/table.csv?s=^TNX&ignore=.csv"&a=04&b=15&c=2003";

A nossa aplicação verifica sempre qual a última data que existe na BD e só apanha o que é necessário (tanto para este como sempre que é possível)

Informação relativa às relações cambiais e preço do ouro:

Encontramos a informação relativa a este indicador no seguinte url:

"http://www.bportugal.pt/rates/cambtx/cambdia.csv";

Este ficheiro contém as relações cambiais desde finais dos anos 90 entre muitas moedas.

As primeiras 8 linhas contêm informação diversa que não nos interessa.

A partir da 9ª linha temos:

Data: 1ª coluna

Relação euro-pound: 9ª coluna

Relação euro-yen: 15ª coluna

Relação euro-dollar: 29ª coluna

Preço do ouro: 34ª coluna

O preço do ouro está em euros e é o custo por 1 onça troy. Uma onça troy são 31,103 gramas.

Informação relativa ao preço do petróleo:

Infelizmente não tivemos tanta sorte como nos outros indicadores em relação à informação histórica relativa ao preço do petróleo. O melhor que encontramos foi um ficheiro XLS com uma série de folhas e numa delas o histórico actualizado do petróleo. Passámos essa informação para o ficheiro de inicialização “brentbarril.txt” que descrevemos acima.

Histórico dos títulos e índices:

Encontramos o histórico (data, preço de abertura, preço de fecho, mínimo, máximo, volume e valor do fecho ajustado) relativo a um título ou índice no seguinte url:

"<http://ichart+randomBase+.finance.dcn.yahoo.com/table.csv?s=>";

Se for um título concatena-se directamente, se for um índice é preciso por um ^ antes.

Assim:

<http://ichart5.finance.dcn.yahoo.com/table.csv?s='^IXIC'>

Tal como no USA 10 year bound rate também podemos concatenar a data a partir da qual queremos os dados.(ver acima)

A randomBase é um inteiro entre 3 e 13 que serve para fazer distribuição da carga.

Os dados deste url vêm em formato csv (comma separated values) o que facilita bastante o parsing.

Informação relativa aos dividendos e splits:

Encontramos informação relativa aos dividendos e splits:

"[http://finance.yahoo.com/q/hp?s="+ticker+"&g=v&z=1000](http://finance.yahoo.com/q/hp?s=)";

Em que ticker é o código da empresa.

Estes dados não vêm em formato csv. Foram precisas expressões regulares para processar os dados.

Informação financeira de uma empresa:

Acabámos por não considerar a informação financeira no nosso modelo.

Se o fizéssemos seriam estes os urls:

Informação financeira por quarters:

<http://finance.yahoo.com/q/is?s=TICKER>

Informação financeira anual:

<http://finance.yahoo.com/q/is?s=TICKER&annual>

Anexo 2 - Dicionário de dados

Dimensão tempo (tabela time)

Date	Data (chave primária da tabela time)
dayOfMonth	Dia do mês (1-31)
month	Mês (1-12)
Year	Ano
dayOfYear	Dia do ano (1-365)
weekInYear	Número da semana no ano (1-52)
semester	Número do semestre (1-2)
trimester	Número do trimestre (1-4)
isWeekend	Fim de semana (1,0)

Sectores

Cod_Sector	chave primária do sector
Desc_sector	Descrição do sector

Industrias

cod_industry	chave primária da tabela industry
Desc_industry	Descrição da industria
cod_sector	Chave externa para a tabela sectors

Empresas (títulos)

cod_company	Chave primária da tabela companies
name	Nome da empresa
start_date	Data do primeiro registo de cotação
num_employees	Numero de empregados da empresa
Cod_industry	Chave externa da tabela industry

Moedas (currencies)

cod_currency	Chave primária da tabela currency
Desc_currency	Nome da moeda

Países (countries)

cod_country	Chave primária da tabela countries
desc_country	Nome do país
cod_currency	Chave externa para a tabela currency

Feriados (holidays)

cod_country	Chave externa da tabela countries
date	Data do feriado

Índices (indexes)

cod_index	Chave primária da tabela indexs
-----------	---------------------------------

desc_index	Descrição do índice
start_date	Data do primeiro registo do índice
num_companies	Número de empresas incluídas no índice
cod_country	Chave externa da tabela countries

Índice do Título

cod_company	Chave externa da tabela companies
cod_index	Chave externa da tabela indexes

Cotações da bolsa

cod_company	Chave externa da tabela companies
Date	Data da cotação
opening	Valor na abertura da sessão
max_day	Máximo da sessão
min_day	Mínimo da sessão
closing	Valor no fecho da sessão
Volume	Volume de títulos transacionados
adjusted_closing	Valor ajustado para splits e merges no fecho da sessão

Dividendos

cod_company	Chave externa da tabela companies
date	Data do dividendo
value	Valor por acção do dividendo

Fusões e divisões

cod_company	Chave externa da tabela companies
Date	Data da fusão ou divisão
titles_before	Títulos antes da fusão ou divisão
titles_after	Títulos correspondentes após fusão ou divisão

Índices (index_quotes)

cod_index	Chave primaria dos índices
date	Data a que correspondem os valores do índice
opening	Valor do índice na abertura
max_day	Valor máximo do dia
min_day	Valor mínimo do dia
closing	Valor no fecho
volume	Volume transaccionado dos títulos pertencentes ao índice
Adjusted_closing	Valor de fecho ajustado

Códigos dos indicadores mundiais (world_indicator_codes)

cod_world_indicator	Código do indicador mundial
desc_world_indicator	Descrição do indicador mundial

Valores dos indicadores mundiais (world_quotes)

cod_world_indicator	Chave primaria dos indicadores mundiais
date	Data do valor do indicador mundial
value	Valor do indicador mundial

Padrões de candle stick (candle_stick_patterns)

cod_candle_pattern	Chave primaria do padrão de candle stick
desc_candle_pattern	Descrição do padrão de candle stick

Indicadores para as cotações da bolsa (stock_indicators)

cod_company	Código do título (chave primária)
Date	Data da cotação (chave primaria)
DifEMA_VAL_5	Diferença à média exponencial de 5 dias
DifEMA_VAL_20	Diferença à média exponencial de 20 dias
DifEMA_VAL_50	Diferença à média exponencial de 50 dias
DifEMA_VAL_200	Diferença à média exponencial de 200 dias
DifWMA_VAL_5	Diferença à média pesada de 5 dias
DifWMA_VAL_20	Diferença à média pesada de 20 dias
DifWMA_VAL_50	Diferença à média pesada de 50 dias
DifWMA_VAL_200	Diferença à média pesada de 200 dias
DifMA_VAL_5	Diferença à média simples de 5 dias
DifMA_VAL_20	Diferença à média simples de 20 dias
DifMA_VAL_50	Diferença à média simples de 50 dias
DifMA_VAL_200	Diferença à média simples de 200 dias
SUPPORT_VAL_200	Valor de suporte calculado para 200 dias
SupportStrength	Força do suporte, nº de encontros com o suporte
DifSupport	Diferença da cotação de fecho ao valor do suporte
RESISTENCE_VAL_200	Valor da resistência calculada para 200 dias
ResistanceStrength	Força da resistência, nº de encontros com a resistência
DifResistance	Diferença da cotação de fecho ao valor da resistência
RSI_VAL_14	Valor do RSI a 14 dias
MACD_VAL_12_26	Valor do MACD a 12(intervalo curto) e 26 dias(intervalo longo)
SIGNAL_MACD_VAL_12_26	Valor do sinal do MACD
FastStochK_5	K do indicador estocástico rápido
FastStochD_5	D do indicador estocástico rápido
SlowStochD_5	D do indicador estocástico lento
DifMA_VOL_50	Diferença do volume á média de 50 dias
ROC_12	Indicador Rate of Change a 12 dias
STDEV_RATIO_50	Desvio padrão sobre o valor médio a 50 dias
codCandlePattern	Código do padrão de candle stick
daysAfterCandlePattern	Dias após o ultimo padrão de candle stick

PASTEVOL_VAL_1	Evolução de 1 dia no passado até á data de hoje
PASTEVOL_VAL_3	Evolução de 3 dias no passado até á data de hoje
PASTEVOL_VAL_7	Evolução de 7 dias no passado até á data de hoje
PASTEVOL_VAL_15	Evolução de 15 dias no passado até á data de hoje
PASTEVOL_VAL_30	Evolução de 30 dias no passado até á data de hoje
PASTEVOL_VAL_60	Evolução de 60 dias no passado até á data de hoje
EVOL_VAL_1	Evolução 1 dia após o presente (coluna de objectivo)
EVOL_VAL_3	Evolução 3 dia após o presente (coluna de objectivo)
EVOL_VAL_7	Evolução 7 dia após o presente (coluna de objectivo)
EVOL_VAL_15	Evolução 15 dia após o presente (coluna de objectivo)
EVOL_VAL_30	Evolução 30 dia após o presente (coluna de objectivo)
EVOL_VAL_60	Evolução 60 dia após o presente (coluna de objectivo)

Indicadores para os índices (index_indicators)

cod_index	Código do índice (chave primária)
Date	Data do valor do índice (chave primaria)
DifEMA_VAL_5	Diferença à média exponencial de 5 dias
DifEMA_VAL_20	Diferença à média exponencial de 20 dias
DifEMA_VAL_50	Diferença à média exponencial de 50 dias
DifEMA_VAL_200	Diferença à média exponencial de 200 dias
DifWMA_VAL_5	Diferença à média pesada de 5 dias
DifWMA_VAL_20	Diferença à média pesada de 20 dias
DifWMA_VAL_50	Diferença à média pesada de 50 dias
DifWMA_VAL_200	Diferença à média pesada de 200 dias
DifMA_VAL_5	Diferença à média simples de 5 dias
DifMA_VAL_20	Diferença à média simples de 20 dias
DifMA_VAL_50	Diferença à média simples de 50 dias
DifMA_VAL_200	Diferença à média simples de 200 dias
SUPPORT_VAL_200	Valor de suporte calculado para 200 dias
SupportStrength	Força do suporte, nº de encontros com o suporte
DifSupport	Diferença da cotação de fecho ao valor do suporte
RESISTENCE_VAL_200	Valor da resistência calculada para 200 dias
ResistanceStrength	Força da resistência, nº de encontros com a resistência
DifResistance	Diferença da cotação de fecho ao valor da resistência
RSI_VAL_14	Valor do RSI a 14 dias
MACD_VAL_12_26	Valor do MACD a 12(intervalo curto) e 26 dias(intervalo longo)
SIGNAL_MACD_VAL_12_26	Valor do sinal do MACD
FastStochK_5	K do indicador estocástico rápido
FastStochD_5	D do indicador estocástico rápido
SlowStochD_5	D do indicador estocástico lento
DifMA_VOL_50	Diferença do volume á média de 50 dias
ROC_12	Indicador Rate of Change a 12 dias
codCandlePattern	Código do padrão de candle stick

daysAfterCandlePattern	Dias após o ultimo padrão de candle stick
EVOL_VAL_1	Evolução 1 dia após o presente (coluna de objectivo)
EVOL_VAL_3	Evolução 3 dia após o presente (coluna de objectivo)
EVOL_VAL_7	Evolução 7 dia após o presente (coluna de objectivo)
EVOL_VAL_15	Evolução 15 dia após o presente (coluna de objectivo)
EVOL_VAL_30	Evolução 30 dia após o presente (coluna de objectivo)
EVOL_VAL_60	Evolução 60 dia após o presente (coluna de objectivo)

Indicadores mundiais (world_indicators)

cod_world_indicator	Código do indicador mundial (chave primária)
Date	Data do indicador mundial (chave primaria)
DifWMA_VAL_5	Diferença à média pesada de 5 dias
DifWMA_VAL_20	Diferença à média pesada de 20 dias
DifWMA_VAL_50	Diferença à média pesada de 50 dias
DifWMA_VAL_200	Diferença à média pesada de 200 dias
DifMA_VAL_5	Diferença à média simples de 5 dias
DifMA_VAL_20	Diferença à média simples de 20 dias
DifMA_VAL_50	Diferença à média simples de 50 dias
DifMA_VAL_200	Diferença à média simples de 200 dias
MACD_VAL_12_26	Valor do MACD a 12(intervalo curto) e 26 dias(intervalo longo)
SIGNAL_MACD_VAL_12_26	Valor do sinal do MACD
EVOL_VAL_1	Evolução 1 dia após o presente (coluna de objectivo)
EVOL_VAL_3	Evolução 3 dia após o presente (coluna de objectivo)
EVOL_VAL_7	Evolução 7 dia após o presente (coluna de objectivo)
EVOL_VAL_15	Evolução 15 dia após o presente (coluna de objectivo)
EVOL_VAL_30	Evolução 30 dia após o presente (coluna de objectivo)
EVOL_VAL_60	Evolução 60 dia após o presente (coluna de objectivo)

Anexo 3 – Figuras para Análise Gráfica

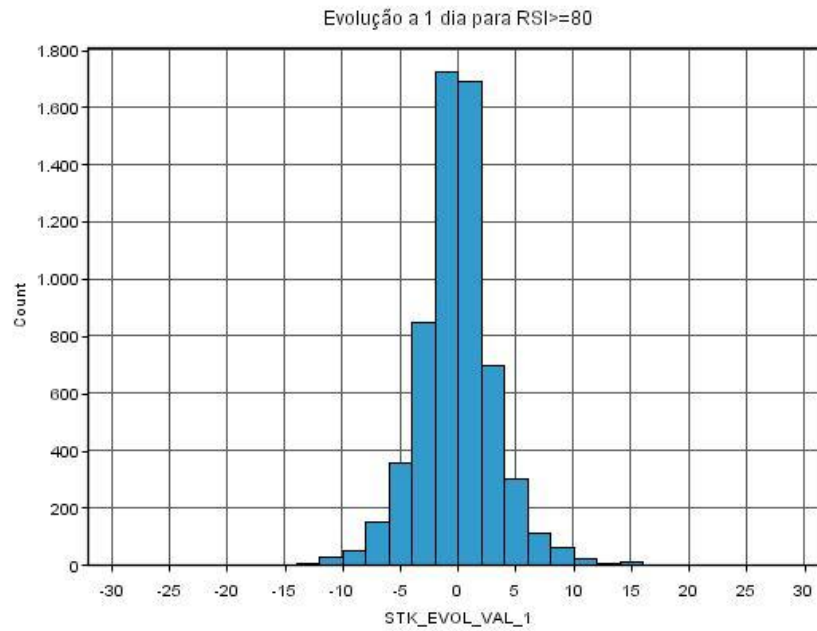


Figura 26: evolução das cotações do índice NASDAQ a 1 dia com RSI ≥ 80

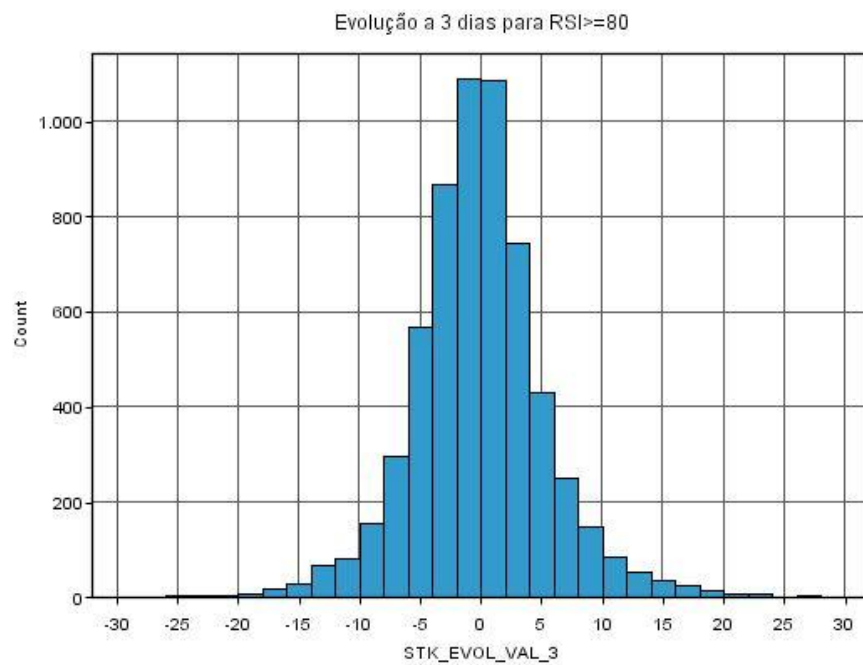


Figura 27: evolução das cotações do índice NASDAQ a 3 dias com RSI ≥ 80

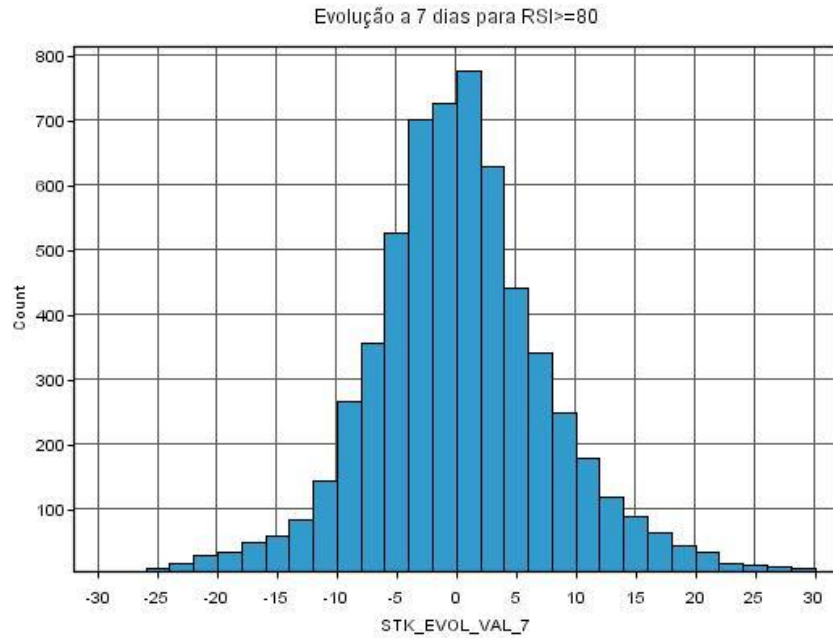


Figura 3: evolução das cotações do índice NASDAQ a 7 dias com $RSI \geq 80$

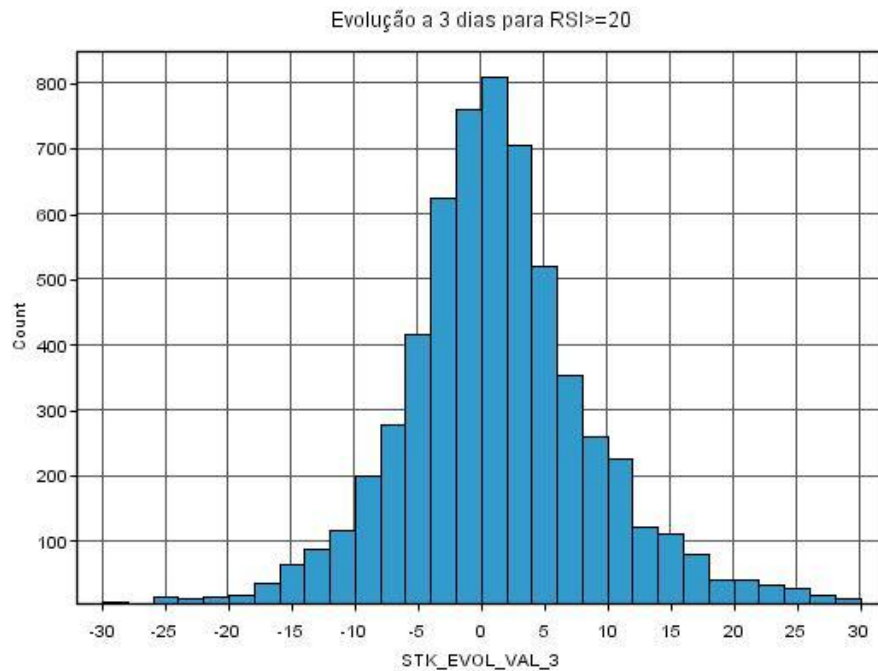


Figura 28: evolução das cotações do índice NASDAQ a 3 dias com $RSI \leq 20$

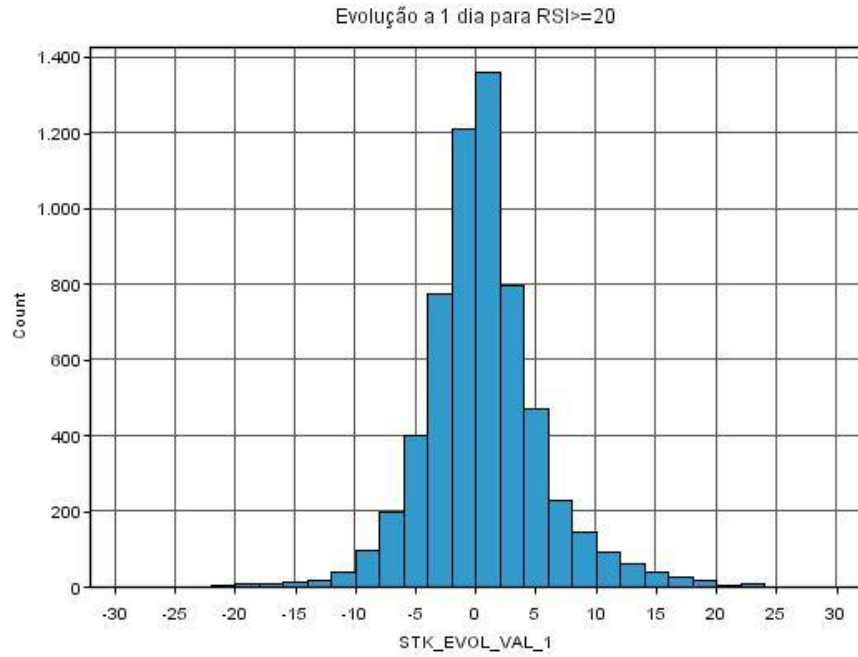


Figura 29: evolução das cotações do índice NASDAQ a 1 dia com $RSI \leq 20$

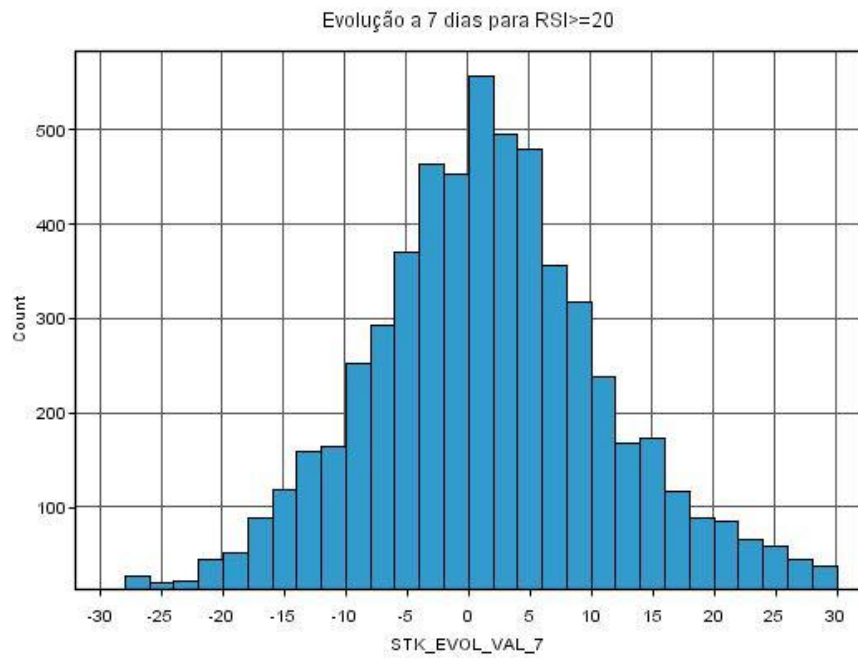


Figura 30: evolução das cotações do índice NASDAQ a 7 dias com $RSI \leq 20$

Anexo 4 – Análise dos resultados dos modelos de Data Mining

Quadro 5: Precisão global dos modelos de Data Mining para o índice NASDAQ face a diferentes níveis de confiança mínimos.

	Confiança	Correctos	Total	Percentagem	Ganho de confiança face ao aleatório
Modelo a 1 dia	0.50	28328	51574	54.93%	4.93%
	0.60	5219	10057	51.89%	1.89%
	0.70	388	699	55.51%	5.51%
	0.80	5	9	55.56%	5.56%
	0.90	0	0	0.00%	-
Modelo a 7 dias	0.50	26037	51574	50.48%	0.48%
	0.60	24022	47426	50.65%	0.65%
	0.70	15077	29958	50.33%	0.33%
	0.80	4520	8935	50.59%	0.59%
	0.90	532	1035	51.40%	1.40%
Modelo a 30 dias	0.50	24997	51574	48.47%	-1.53%
	0.60	24482	50555	48.43%	-1.57%
	0.70	22311	45702	48.82%	-1.18%
	0.80	17286	35311	48.95%	-1.05%
	0.90	7605	15073	50.45%	0.45%
Modelo a 3 dias	0.50	26037	51574	50.48%	0.48%
	0.60	16306	31885	51.14%	1.14%
	0.70	3984	7823	50.93%	0.93%
	0.80	521	1052	49.52%	-0.48%
	0.90	0	0	0.00%	-
Modelo a 15 dias	0.50	26016	51574	50.44%	0.44%
	0.60	25257	49958	50.56%	0.56%
	0.70	21509	42780	50.28%	0.28%
	0.80	12049	23625	51.00%	1.00%
	0.90	3921	7594	51.63%	1.63%
Modelo a 60 dias	0.50	27227	51574	52.79%	2.79%
	0.60	26815	50662	52.93%	2.93%
	0.70	25096	47331	53.02%	3.02%
	0.80	21201	39581	53.56%	3.56%
	0.90	11263	20642	54.56%	4.56%

10 Melhores Modelos

Quadro 6: Estatísticas dos modelos a 1 Dia para um nível de confiança mínimo de 65%

	Código do Título	Precisão	Número de correctos/Número total
1	WFMI	83,33	(5/6)
2	SSCC	81,08	(30/37)
3	GNTX	79,59	(39/49)
4	ROST	78,26	(18/23)
5	SNPS	76,19	(16/21)
6	IVGN	73,47	(72/98)
7	ERTS	73,02	(46/63)
8	DLTR	72,73	(16/22)
9	COST	72,73	(16/22)
10	DELL	72,73	(32/44)

Quadro 7: Estatísticas dos modelos a 3 Dias para um nível de confiança mínimo de 65%

	Código do Título	Precisão	Número de correctos/Número total
1	CMCS	72,2	(65/90)
2	MRVL	66,67	(6/9)
3	IACI	65,22	(105/161)
4	ERTS	63,56	(150/236)
5	NVLS	61,54	(32/52)
6	SPLS	60,33	(111/184)
7	TLAB	60,29	(41/68)
8	PSFT	57,40	(128/223)
9	AAPL	57,09	(141/247)
10	DLTR	57,06	(97/170)

Quadro 8: Estatísticas dos modelos a 7 Dias para um nível de confiança mínimo de 75%

	Código do Título	Precisão	Número de correctos/Número total
1	VRSN	81,82	(9/11)
2	AMZN	76,27	(45/59)
3	IACI	72,00	(90/125)
4	AAPL	71,43	(40/56)
5	YHOO	71,23	(52/73)
6	CMCS	66,6	(20/30)
7	PDCO	64,04	(57/89)
8	CTXS	62,62	(67/107)
9	ERTS	61,05	(174/285)
10	CEPH	60,33	(73/121)

Quadro 9: Estatísticas dos modelos a 15 Dias para um nível de confiança mínimo de 80%

	Código do Título	Precisão	Número de correctos/Número total
1	CECO	95,00	(19/20)
2	GRMN	71,30	(82/115)
3	GNTX	68,00	(68/100)
4	BMET	66,16	(131/198)
5	EBAY	66,16	(303/458)
6	CSCO	64,77	(57/88)
7	JNPR	62,30	(157/252)
8	CPWR	62,02	(80/129)
9	BIIB	61,86	(146/236)
10	PSFT	61,62	(114/185)

Quadro 10: Estatísticas dos modelos a 30 Dias para um nível de confiança mínimo de 80%

	Código do Título	Precisão	Número de correctos/Número total
1	GRMN	87,76	(86/98)
2	XRAY	76,79	(182/237)
3	FHCC	76,34	(100/131)
4	NXTL	76,05	(289/380)
5	AMZN	75,24	(319/424)
6	CTAS	73,49	(352/479)
7	EBAY	72,61	(342/471)
8	AMGN	71,10	(251/353)
9	JNPR	69,50	(303/436)
10	CHIR	68,00	(153/225)

Quadro 11: Estatísticas dos modelos a 60 Dias para um nível de confiança mínimo de 80%

	Código do Título	Precisão	Número de correctos/Número total
1	CHRW	100,0	(43/43)
2	CMVT	88,98	(210/236)
3	SBUX	88,39	(99/112)
4	APOL	84,26	(364/432)
5	PTEN	80,91	(195/241)
6	FHCC	80,58	(390/484)
7	NXTL	79,27	(325/410)
8	MCHP	74,11	(292/394)
9	MLNM	72,79	(305/419)
10	AMGN	72,03	(273/379)

Anexo 5 – Ficheiros diversos

Ficheiros utilizados no processo de avaliação dos modelos de Data Mining

ComparativoModelos2002.xls
ComparativoModelos2003.xls

Ficheiros de Stream do Clementine®

DM_Stream.str (stream manipulado pelos scripts)
DM_Model.clm (script de manipulação do stream DM_Stream.str)
BolsaStream.str (Stream manual)

Ficheiros para apoio na criação da tabela da Data Mining

dataMining.cs

Ficheiros de criação da base de Dados

createDB.sql