Subject Name: Principles and Practice of Data Mining
Subject No.: 32130
Student No.: 91093688
Email: gerard@aba.com.au
Phone No (H): (02) 94208894 **(W):** (02) 9369 4069
Assignment No.: 3

TOPIC: DATA MINING PROPOSAL – DETECTION OF INSIDER TRADING

Abstract. This document comprises a proposal for a data mining project focused on the detection of insider trading in a major stock exchange. The proposal is based on the combination of textual news feed information and stock discussion transcripts with trade price and volume history. The addition of trading entity relationship data is introduced to provide a rich data mining environment in which to discover the rules that govern insider trading. The document initially discusses existing insider trading detection systems and data mining techniques before focusing on the data required, potential output and the recommended tools for the project, SPSS Lexiquest and SPSS Clementine.

I HAVE CAREFULLY READ, UNDERSTOOD, AND HAVE TAKEN INTO ACCOUNT ALL THE REQUIREMENTS AND GUIDELINES FOR ASSESSMENT AND REFERENCING IN THE SUBJECT OUTLINE. I AFFIRM THAT THIS ASSIGNMENT IS MY OWN WORK; THAT IT HAS NOT BEEN PREVIOUSLY SUBMITTED FOR ASSESSMENT; THAT ALL MATERIAL WHICH IS QUOTED IS ACCURATELY INDICATED AS SUCH; AND THAT I HAVE ACKNOWLEDGED ALL SOURCES USED FULLY AND ACCURATELY ACCORDING TO THE REQUIREMENTS. I AM FULLY AWARE THAT FAILURE TO COMPLY WITH THESE REQUIREMENTS IS A FORM OF CHEATING AND COULD RESULT IN DISCIPLINARY ACTION.

Table of Contents and Figures

Table of Contents and Figures 2
1. Introduction 2
2. Aims, Objectives and Possible Outcomes 3
2.1 Aims 3
2.2 Objectives 3
2.3 Possible Outcomes 3
3. Background 3
3.1 Current Systems 3
3.1.1 MonITARS (Monitoring Insider Trading and Regulatory Surveillance) 3
3.1.2 National Association of Securities Dealers (NASD) Advanced-Detection System (ADS) ... 4
3.1.3 NASD Securities Observation, News Analysis and Regulation (SONAR) 4
3.1.4 SOMA (Surveillance of Market Activity) 5
3.2 known problems 5
4. Data mining scenario and methodology based on CRISP-DM 6
4.1 Business Understanding 6
4.1.1 Business objectives 6
4.1.2Assessment of situation 6
4.1.3 Data mining goals 6
4.1.4 Project plan 7
4.2 Understanding the data 8
4.2.1 Source Data 8
4.2.2 Output 9
4.3 Data mining tools 10
5. Conclusion 10
Appendix A. Bibliography 11

Figure 1. ADS Break Browser (Kirkland, 1999) 4
Figure 2. Sample news feed from IRESS, an Australian on-line trading system 9
Figure 3. How SPSS Lexiquest processes documents into usable data (SPSS, 2003) 10

1. Introduction

This document is a project proposal for the creation of data mining project to assist in the detection of insider trading. The scope also includes more general market movement predictions based on existing market data and external influences. It first looks at the aims and objectives to be focused on when developing an insider trading detection system before focussing on the possible outcomes of the development of such a system. The background section of the document examines current technologies being used in the area and systems currently in use in some of the world's major stock exchanges.

The data mining scenario has been based around the CRISP-DM methodology, focussing on the business understanding, data understanding and data preparation elements of the CRISP-DM process. The potential results will be examined from a hypothetical perspective, highlighting methods of attaining the results.

2. Aims, Objectives and Possible Outcomes

2.1 AIMS

The key aim of producing an insider trading detection system is to increase the probability that someone trading with inside information will be detected and caught. As stock exchanges stand, the sheer volume of data that needs to be analysed to detect insider trades is massive, upwards of 2 million transactions per day (Kirkland, 1999),and can no longer be done using manual processes as has been done in the past.
Using data mining techniques, it will be possible on a daily basis to flag a manageable set of trades that potentially could be insider trades. These trades can then be manually investigated by an investigation team, removing the huge load of manual number crunching tasks away from the investigation teams and letting them get on with their job of determining which of the trades detected by the system are fraudulent.

2.2 OBJECTIVES

The system's key objective will be to produce a daily set of results in a clear human readable form that will show detected trade patterns that are potential insider trades. The data available to the system to produce this result set consists of all the trade information from the exchange, a database of financial news in textual format as well as transcripts of discussions on securities that are traded at the stock exchange. Through the use of text mining technologies, the transcripts and news feeds can be linked to each day's trades to try and determine insider trading activity.

2.3 POSSIBLE OUTCOMES

The targeted outcome for the system is to improve the detection of illegal insider trading. This can be accomplished by increasing the likelihood that cases detected are insider trading, reducing the number of cases the investigation team need to thoroughly examine.
The key limiting factor in this development will be cost. As several major exchanges are already successfully using data mining techniques for this task, it is reasonable to expect that if the development is managed properly, the project will be a success from a functionality standpoint. The cost of building the system and the available budget will be the limiting factors as to how successful the results of the project will be.
Improving the detection system for insider trading would have the added benefit of significantly reducing insider trading by deterring those who otherwise may be tempted from taking the risk of being caught. Investment in this area also has an important signalling effect on the general public as well as investors where the perception of insider trading occurring at the expense of other investors could be reduced.

3. Background

3.1 CURRENT SYSTEMS

3.1.1 MonITARS (Monitoring Insider Trading and Regulatory Surveillance)

Accoding to Finley (Finley, 2000), the London Stock Exchange uses the MonITARS system developed by Searchspace, a self proclaimed leading provider of intelligent enterprise systems, is designed to "detect suspicious trades within the vast amount of data that electronically passes through the exchange each day" (Finley, 2000). This is achieved through the use of Neural Network technology.

According to Jan Tellick, Searchspace's marketing director, "The system can even detect dealing rings where several individuals or one individual with several accounts trade across all their accounts for the purpose of insider dealing or market manipulation." (Finley, 2000). The MonITARS system doesn't only detect insider trading, but also front running, market manipulation via wash trades, marking the close, influencing the open, phantom orders and unusual price/volumes (Searchspace, 2003). This is achieved through the use of an internal profiling database that maintains data surrounding each of the investors and their known relationships other investors and securities.

3.1.2 National Association of Securities Dealers (NASD) Advanced-Detection System (ADS)
The ADS was designed to identify patterns and practices of behaviour that could be of regulatory interest (Kirkland, 1999). The system processes over 2 million transactions per day generating a list of around 10,000 potentially illegal transactions. The system is based around a pattern matching engine which is based on patterns or rules determined by the data mining component of the system.
The data mining component of ADS uses 'parallel and scalable decision tree and association rule implementations that can be used to discover new rules reflecting changing behaviours in the marketplace' (Kirkland, 1999). The association rule algorithm for the system is implemented in C++ using an Oracle 8 relational database for data management. Various custom filters have been implemented to highlight rules of particular interest.
The decision tree algorithm in ADS is used to identify patterns with respect to a particular attribute. The combination of association rule and decision tree data mining techniques are used to produce a set of around 10-20 rules each day. These rules are then examined by a group of analysts and effective rules are added to the pattern matching engine for illegal activity detection.

Figure 1. ADS Break Browser (Kirkland, 1999)

The break browser (figure 1) is the main window in ADS used for reviewing and analysing the various potential illegal transactions generated by the system. The top half of the window displays a set of standard attributes and can be used to filter and sort the analyst's set of open potential regulatory breaks. The lower section displays a detailed description of the selected break. This is one of a number of tools that allow the users of ADS to firstly investigate a break then substantiate it and build a case for regulatory action. (Kirkland, 1999)

3.1.3 NASD Securities Observation, News Analysis and Regulation (SONAR)
SONAR has been the next phase in the NASDs attack on illegal trading activities. According to KDNuggets (2003), SONAR has been in action since December 2001. It processes around 6.5 million

news wire stories and 350,000 SEC filings from corporations each year. These are combined with the price and volume models for 25,000 securities each day to produce between 50 and 60 alerts for regulatory analysts to investigate.
SONAR generates only one third of the alerts compared with ADS, however through the combined use of text mining with other data mining techniques, it is three times more accurate in generating 'real' alerts. It also presents the data in a more concise and human readable form, reducing the time required to review a case by around 50%. According to KDNuggets (2003), this saving has translated into a redeployment of 9 of their 30 analysts from dealing with break detection to later stages of review. SONAR uses a combination of natural language processing (NLP), combined with text mining and traditional data mining techniques to provide significant cost savings to the NASD while also improving the effectiveness of the regulatory processes the NASD are required to monitor.

3.1.4 SOMA (Surveillance of Market Activity)

SOMA is the system employed by the Australian Stock Exchange (ASX) to monitor illegal market activity. Although it is over 7 years old, SOMA according to the ASX is still in use today.

> 'SOMA' alerts ASX when there is an occurrence in the trading of a security that is outside pre-set values of any of a number of parameters. In addition, the ASX has developed a computer program called 'SEATSCAM' which allows it to replay the whole market in a stock at a sufficient pace to allow an analyst to examine in detail the trading that has occurred. (Grabosky and Braithwaite, 1993)

From the research, SOMA does not use any data mining technology to detect illegal trading. It rather relies on filtering techniques to give analysts a better way to examine trading patterns. SOMA is an out of date system and has been mentioned here to show the level of automation in the Australian Stock Exchange to give an idea of the current state of insider trading regulation in Australia.

3.2 KNOWN PROBLEMS

Unfortunately, given the nature of these systems, little information is available as to their internal workings or known problems. If the underlying fundamentals of these systems are open for public scrutiny, people trying to hide insider trading and other fraudulent activities could potentially have the know how to circumvent the current systems. As a result of this, almost no details of system implementations can be found through research, other than the basic source data being used, data mining techniques applied in the systems and names of the companies producing the systems. There is also an element of secrecy similar to that of the formula for Coca-cola where there is a lot of competition in this area of development with many of the world's smaller exchanges yet to implement systems; existing suppliers are guarding their trade secrets and data mining techniques that comprise a competitive advantage very carefully.

4. Data mining scenario and methodology based on CRISP-DM

4.1 BUSINESS UNDERSTANDING

4.1.1 Business objectives

The business focus for this proposal is to provide detection facilities to detect illegal trading activity for major stock exchanges. Financial data has been collected for all major stock exchanges for a number of years. This includes historical transaction data, financial news records and transcripts of discussions about stocks being traded. It is proposed that this data can be used through data mining techniques to provide additional benefits to the stock exchange in the form of financial predictions and fraud detection of insider trading.

The key business objective of this proposal is to provide the facility to detect illegal insider trading through the use of data mining technology. The success criterion for this project is the discovery of identifiable patterns that indicate illegal insider trading.

4.1.2Assessment of situation

There are a number of data elements that will need to be combined in order to achieve the business objectives through the data mining process. The base data, as used in all existing insider trading detection systems involves the use of daily stock market trade and volume information to determine purchasing and selling patterns. This information is readily available and has been collected by all major exchanges for many years. Combining this information with security announcements, daily stock market news feeds and transcripts of stock discussions will provide significantly stronger results. This has been demonstrated by the SONAR system employed by the NASD providing up to three times the accuracy over the use of stock trade volume data alone.

Another data element that should be added to the data mining process is information about traders and their relationships with various corporations. Data should be collected regarding board members, senior management and their families with links to the companies they are responsible for.

Existing systems use a wide range of data mining methods to produce their results. The MonITARS system employed by the London stock exchange is based on neural networks whereas both the systems employed by the NASD in the United States of America use a combination of decision tree and association rule mining techniques. Given the legal nature of insider trading, it seems logical to use association rule mining to produce more effective results. Once a suspect trade is determined, analysts must then determine why it may be an illegal trade and then go about building a case for the authorities to pursue. A neural network would be effective for identification but the analyst may not be able to determine for what reasons the neural network identified the trade. Neural networks provide only a black box for identifying patterns, but no way of seeing inside that black box to determine the principles that led to the trade being detected.

In order to make good use of the news feeds and discussion transcripts, a combination of natural language processing and text mining will be required. From looking at the systems currently in use, this combination of data combined with natural language processing (NLP), text mining and association rule mining, should provide a potent combination for mining insider trading patterns.

4.1.3 Data mining goals

The key goal of this project is to provide a set of identifiable patterns that indicate when insider trading has potentially occurred, allowing analysts to then formulate cases around the highlighted trades. These patterns will take the form of rules that highlight when a certain set of conditions are

met, it is likely that insider trading has occurred. Due to the combination of text mining and data mining techniques required, the combined data mining toolkit SAS Enterprise Miner will be used in combination with SAS Text Miner.

4.1.4 Project plan

A basic project plan is shown below to give a guide to the timing of the project. The project plan highlights the various phases of the process, as described by the CRISP-DM methodology along with the estimated man hours required to complete them. Based on the time estimated and the use of a single data mining expert, the initial data mining project could be complete in just under 3 months. This rapid turnaround would allow the exchange to put this new technology to the test as soon as possible, increasing the potential to restrict illegal activity sooner rather than later.

Task	Person Hours
1. Collect initial data	
1.1. Describe data	16
1.2. Explore data	24
1.3. Verify data quality	8
2. Data preparation	
2.1. Select data	16
2.2. Clean data	40
2.3. Construct data	40
2.4. Integrate data	40
2.5. Format data	24
3. Modeling	
3.1. Select modeling technique	16
3.2. Generate test design	40
3.3. Build model	16
3.4. Assess model	24
4. Evaluation	
4.1. Evaluate results	8
4.2. Review process	24
4.3. Determine next steps	8
5. Deployment	
5.1. Plan deployment	40
5.2. Plan monitoring and maintenance	32
5.3. Produce final report	24
5.4. Review project	24

Table 1. Brief project plan

4.2 UNDERSTANDING THE DATA

4.2.1 Source Data

The main source of data will be stock market trade data. The format for this data is outlined in the following table.

Attribute	**Type**	**Definition**
DateTime	Interval	The date and time the transaction occurred
Security	Nominal	Security code of the security being traded (BHP,ANZ,QBE)
Seller	Nominal	CHESS code of the seller
Buyer	Nominal	CHESS code of the buyer
Price	Ratio	Share price at which the trade was completed
Volume	Ratio	Quantity of shares exchanged
BuyBroker	Nominal	CHESS code of the buyer's broker
SellBroker	Nominal	CHESS code of the sellers broker
Legitimate	Binary	True or false: Used for training data and as the production field for live data. Distinguishes whether the trade was an insider trade

Table 2. Trade data structure

A database of trading entities with relationships to securities will be maintained in the following format

Attribute	**Type**	**Definition**
CHESS Code	Nominal	The CHESS code of the entity (company or individual)
Security Code	Nominal	Security code of the security the entity is related to.
Relationship	Nominal	A set of categories of relationship (Board Member, CEO etc…)

Table 3. Relationship data structure

The news feed and security discussion data will take the form of written English text that will be preprocessed with a natural language processor (NLP) before being included in the data mining process through the use of text mining. News feeds will be stored in the following format.

Attribute	**Type**	**Definition**
DateTime	Interval	The date and time the transaction occurred
Security Code	Nominal	Security code of the security the news item relates to
Source	Nominal	The source of the news item (eg: ASX, Reuters)
Description	Text	Textual 1 line description of the news item
News	Text	The full text of the news feed item

Table 4. News feed data structure

Figure 2. Sample news feed from IRESS, an Australian on-line trading system

Security discussion transcripts will be in a similar format to news feeds and will be integrated into the system using the same combination of NLP and text mining.

Attribute	Type	Definition
DateTime	Interval	The date and time the transaction occurred
Security Code	Nominal	Security code of the security the news item relates to
Source	Nominal	The source of the discussion
Description	Text	Textual 1 line description of the news item
Transcript	Text	The full text of the discussion regarding the security

Table 5. News feed data structure

The date and time of each type of item, combined with its security code, will be used to correlate the various pieces of data and information into a coherent data set that can be successfully mined to produce the desired results.

4.2.2 Output

The combined data will allow the data mining algorithms to cater for events such as a sudden price increase after a news item covering a potential takeover of a security. A trade by a related entity prior to the public announcement would likely be suspect. This is one of the more obvious insider trading examples that current manual regulations examine, the data mining process would allow far more subtle situations to be detected and highlighted for further analysis.

The output from the project will be rule sets that allow the determination of potentially illegal trades. A fictitious example rule to demonstrate the look and feel of the output is shown below.

Prediction: **Legitamate** is **<u>False</u>**
Conclusive Prediction's probability: **0.924**
Rule(s) explaining why the case is unexpected.

1) *If* **Security** *is* **<u>ANZ</u>**
and **Relationship** *is* **<u>CEO</u>**
Then
Legitimate *is* **<u>False</u>**
Rule's probability: **0.924**
The rule exists in **26** *records.*
Significance Level: Error probability < 0.1

4.3 DATA MINING TOOLS

The data will be combined with the output of the text mining to produce lists of key concepts and their frequencies for textual data. SPSS Lexiquest Mine is the product that will be used for the text mining as not only does it produce the desired output from the textual data but it integrates with SPSS Clementine, a full enterprise data mining suite. This will allow streamlined handling of the pre-processing and integration of the trade and relationship data with the textual data right through to execution of the data mining algorithms and reporting of results.
According to SPSS, Lexiquest mine will perform as follows.

> For each article of text, such as a patent or e-mail, linguistic-based text mining returns an index of concepts as well as the frequency and class of each concept. This distilled information can be combined with other data sources and used with traditional data mining techniques such as clustering, classification and predictive modeling. (SPSS, 2003)

Figure 3. How SPSS Lexiquest processes documents into usable data (SPSS, 2003)

SPSS is one of the leading producers of world class data mining tools with a strong emphasis on return on investment (ROI). It is one of the tools recommended by www.kdnuggets.com, one of the leading data mining information resources on the internet.

5. Conclusion

After thorough research of the data mining systems currently in use around the world, this project will utilize the best of breed data mining techniques on a combination of source data that has been proven effective in two of the largest exchanges in the world. The use of association rule mining will allow the required level of legal transparency in the data mining process to show why certain trades have been singled out as potential insider trades. The use of trading entity relationships will significantly improve the ability to detect minor infringements as has been proven in the MonITARS system in London.
The project will unite some of the best text mining and data mining tools through the use of SPSS Lexiquest and Clementine with complete trade, relationship and textual data surrounding the various securities on the exchange. This will allow us to produce high quality sets of rules and patterns that will greatly improve the detection of insider trading through the use of automated pattern matching techniques. Additionally, the project can be completed within a 3 month time frame allowing rapid return on investment (ROI).

Appendix A. Bibliography

(2002), 'Mantas Launches Broker Surveillance Monitor', *DSstar E-zine* [Online]. Available: http://www.tgc.com/dsstar/02/0611/104344.html [Accessed November 1, 2003].

(2003), 'AI system used for Monitoring NASDAQ for Potential Insider Trading and Fraud', *KDnuggets* [Online]. Available: http://www.kdnuggets.com/news/2003/n18/13i.html [Accessed October 24, 2003].

(2003b), 'Detection of insider trading', *New Zealand Ministry of Ecomonic Development* [Online]. Available: http://www.med.govt.nz/buslt/bus_pol/bus_law/insidertrading/insidertrading-08.html [Accessed November 2, 2003].

(2003c), 'The SEC wants to know: was an executive's sudden sale of 100,000 shares legitimate, or was it insider trading?' *Zeeware* [Online]. Available: http://www.zeeware.com/zw/downloads/caseStudies/CaseStudyInvestigate.pdf [Accessed October 29, 2003].

(2003d), 'Surveillance of trading activity', *Australian Stock Exchange* [Online]. Available: http://www.asx.com.au/about/l3/Surveillance_AA3.shtm [Accessed November 2, 2003].

Abbott, D. W. M., IPhilip; Elder, John FIV (1998) "Evaluation of high-end data mining tools for fraud detection", *The 1998 IEEE International Conference on Systems, Man, and Cybernetics*, Vol. 3 IEEE, PISCATAWAY, NJ, (USA), San Diego, CA, USA, 11-14 October 1998, pp. 2836-2841.

Bolton, R. J. and Hand, D. J. (2001) "Significance tests for patterns in continuous data", *Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on*, pp. 67-74.

Bolton, R. J. and Hand, D. J. (2002) Statistical Fraud Detection: A Review, *Statistical Science,* **17** (3), pp. 235-255.

Cuneo, E. C. (2003), 'Beyond Compliance - Financial companies want to turn regulatory burden into competitive advantage', *InformationWeek* [Online]. Available: http://informationweek.com/story/IWK20030223S0002 [Accessed October 28, 2003].

Finley, M. (2000), 'Smart Methods to Spot Fraud', *Wired News* [Online]. Available: http://www.wired.com/news/business/0,1367,35273,00.html [Accessed November 1, 2003].

Franklin, D. (2002), 'Data Miners - New software instantly connects bits of data that once eluded teams of researchers', *Time Online Edition* [Online]. Available: http://www.time.com/time/globalbusiness/article/0,9171,1101021223-400017,00.html?cnn=yes [Accessed November 1, 2003].

Grabosky, P. and Braithwaite, J. (1993) *Business Regulation and Australia's Future,* Australian Institute of Criminology.

Hand, D. J., Adams, N. M. and Bolton, R. J. (2002) *Pattern Detection and Discovery,* Springer-Verlag Berlin Heidelberg, London, UK.

Jones, A., Kovacich, G. L. and Luzwick, P. G. (2002) Everything You Wanted to Know about Information Warfare but Were Afraid to Ask, Part 1, *Information Systems Security,* **11** (4), pp. 9-20.

Kirkland, J. S. (1999) NASD regulation advanced-detection system (ADS), *AI Magazine,* **20** (1), 1999, pp. 55-67.

Lopez, N. (2002), 'Eyeing Every Trade', *NYSE Magazine* [Online]. Available: http://www.workingwithwords.com/InsideNYSE.may.june.pdf [Accessed November 2, 2003].

Miley, M. (1998) Taking Stock: Oracle-Based Knowledge System Helps Detect Securiti... *Oracle Magazine,* **12** (3), May/Jun 1998, pp. 62-66.

Pyle, D. (1999) Markets and information -- Data mining techniques can definitively measure how much useful information is contained in market data, but using them... *Intelligent Enterprise (United States),* **2** (8), June 1, 1999, pp. 18.

Safer, A. M. (2003) A comparison of two data mining techniques to predict abnormal stock market returns., *Intelligent Data Analysis,* **7** (1), 2003, pp. 3.

Searchspace (2003), 'Industry Sectors - Exchanges', *Searchspace* [Online]. Available: http://www.searchspace.com/industry/exchanges.shtml [Accessed November 2, 2003].

Sirikulvadhana, S. (2002) *Data Mining As A Financial Auditing Tool,* The Swedish School of Economics and Business Administration.

SPSS (2003), 'Combine text mining with data mining to turn your information into business intelligence', [Online]. Available: http://www.spss.com/pdfs/TMCLMINS-0702.pdf [Accessed November 4, 2003].

Westphal, C. and Blaxton, T. (1998) *Data Mining Solutions: Methods and Tools for Solving Real-World Problems,* John Wiley & Sons, Inc.