AN APPLICATION OF DATALOGIC/R KNOWLEDGE DISCOVERY TOOL TO IDENTIFY STRONG PREDICTIVE RULES IN STOCK MARKET DATA

Wojciech Ziarko and Robert Golan
Computer Science Department

University of Regina
Regina, SK, S4S 0A2, Canada
e-mail: ziarko@cs.uregina.ca

Donald Edwards
First Marathon Securities

McCallum Hill Centre
1847 Scarth Street
Regina,SK, S4P-4B3, Canada

Abstract

An application of a methodology for discovering strong probabilistic rules in data is presented. The methodology is based on an extended model of rough sets called variable precision rough sets model incorporated in DATALOGIC/R knowledge discovery tool from Reduct Systems Inc. It has been applied to analyze monthly stock market data collected over a ten year period. The objective of the analysis was to identify dominant relationships among fluctuations of market indicators and stock prices. For the purpose of comparison, both precise and imprecise, strong and weak rules were discovered and evaluated by a domain expert, a stock broker. The evaluation revealed that the strong rules (supported by many cases) essentially confirm the expert's experiences whereas weak rules are often difficult to interpret. This suggests the use of rule strength as the primary criteria for the selection of potentially useful predictive rules.

1. INTRODUCTION

A new breed of computer programmers have emerged from the investment industry called Quants. Quants use their computer programs to aid brokers and investment managers in their market predictions. This new science originated from the artificial intelligence (AI) push during the early 80's promising an easy way to predict the stock market. It was quickly determined that this was not an easy feat and that a complex interaction between the human experts(brokers) and the computer systems along with good stock market data were needed. Quants have filled the gap to make this interaction work in order to help in building expert systems which would predict the market. The goal of these expert systems is to give a brokerage firm or an investment manager a slight edge over their competitors. This could mean extra profits in the millions of dollars. Details of the success stories are kept confidential so as not to give up their technological edge for market predictions. Nevertheless, there exists quite extensive literature describing attempts to use AI techniques, and in particular neural networks, for predicting stock market variations [1,7,8,9] . The main problem with neural networks, however is the tremendous difficulty in interpreting the results. The neural net approach is essentially a "black box approach" in which no new knowledge regarding the nature of the interactions between the market indicators and the stock market fluctuations is extracted from the market data. Consequently, there is a need to develop methodologies and tools which would help in increasing the degree of understanding of market processes and, at the same time, would allow for relatively accurate predictions. Although statistical methods serve the purpose of acquiring some new knowledge about the data, they do not provide sufficient predictive capability when it comes to problems involving interactions among many interdependent variables with unknown probability distributions. In this context, the methods stemming from the research on knowledge discovery in databases seem to provide a good mix of predictive and knowledge acquisition capabilities for the purpose of market prediction and market data analysis.

In what follows we describe the first results of an on-going research project on the application of knowledge discovery methods based on the variable precision model of rough sets (VPRS) to acquire new knowledge from market data [4,5]. The project has been undertaken in conjunction with First Marathon Securities Ltd. brokerage firm which provided the necessary market expertise in locating the relevant market data, data preprocessing, and verification of results. The computational results have been produced using the beta copy of DATALOGIC/R Version 2.25 of the PC-based system for knowledge discovery in data [6] incorporating elements of the theory of rough sets [2] and of the VPRS model. DATALOGIC/R has been developed by Reduct Systems Inc. of Regina, Canada.

2. BACKGROUND INFORMATION

The stock market is a market for the sale and purchase of stocks and bonds for companies, municipalities, and certificates representing commodities of trade. The stock market plays an important role in any capitalist economy. Companies at times need extra capital to do their business and one way of doing this is by issuing stocks or bonds to investors in exchange for money. The investor is hopefully rewarded by the company with dividends and by stock appreciation. The initial sales of company stocks were made by investment bankers and then the stock exchanges provided the markets with a middle man called the broker. Stock exchanges exist throughout the world. One of the largest is the New York Stock Exchange, which was founded by 24 brokers back in May of 1792. By 1800, 335 companies were registered with the exchange trading mostly bonds. The actual buying and selling of the stocks/bonds is done right on the trade floor of the exchange. These deals are then read into a computer which in turn displays the stock symbol prices world wide on electrical tickers and display devices. The main Canadian stock exchange is the Toronto Stock Exchange(TSE). Technological advancements has made the TSE one of the most modern stock exchanges in the world with such services as Market by Price(MBP). This gives brokers and interested investors a snapshot of the market as it stands at any given moment. The TSE is introducing another first to the investment community when they become the first to implement fully electronic trading. The stock trading will be taken off the exchange floor and put automated computer trading into the hands of the brokers and investors. The TSE is the stock market which is used for analysis in this paper.

3. STRONG RULES DISCOVERY PROBLEM

The problem is to apply knowledge discovery (KD) techniques to identify strong predictive rules from stock and economic data which are a true indication of what happened during a certain time period in the stock market. By strong rules we mean rules reflecting highly repetitive patterns occurring in data. They are not supposed to be necessarily precise or deterministic, they may be associated with fractional probabilities of the predicted outcomes. They should be, however correct or almost correct reflecting real relationships occurring in the economic system. The initial research results on discovery and analysis of such strong rules have been reported by Piatetsky-Shapiro in [3]. The strength of rules can be measured in terms of data objects (data records) satisfying rule conditions. The strong rules are potentially interesting data patterns and likely generally true data regularities. Ideally, if the collected data is a representative and random subset of all feasible combinations of market indicators then it can be proven using probability theory that the stronger the rule the higher the likelihood that the rule represents a true fact, or a close approximation of a true fact about the domain of interest. The discovered strong rules are the subject of further verification by the domain expert, stock broker in our case. The goal of the verification is to prune the "noise" rules i.e. the ones which are clearly in contradiction with broker's experience and whose relative strength follows from the inadequacy of the available data.

4. THE BASICS OF THE VARIABLE PRECISION METHOD OF ROUGH SETS

The original rough sets model as introduced by Pawlak [2] is concerned with the analysis of deterministic data dependencies. In its formalism it does not recognize the presence or absence of non-deterministic relationships, i.e. the ones which may lead to predictive rules with probabilities less than one. In some data sets, however, the available information is not sufficient to produce strong deterministic rules but it may be quite possible to identify strong nondeterministic rules with estimates of decision probabilities. To deal with this problem an extended VPRS model of rough sets was proposed by Ziarko [4,5].

The basic concepts of the VPRS model utilized in the implementation of the DATALOGIC/R system are:

- * approximation space
- * lower BETA-approximation a set where BETA is a real number from the range <0,1>

4.1. APPROXIMATION SPACE

The approximation space has two components:

- * the universe of discourse, or our domain of interest denoted as U.
- * an equivalence relation R partitioning the universe into disjoint classes called elementary sets.

The equivalence relation represents our classification knowledge, i.e. our ability do discern different objects of the universe U. In the stock market problem the universe is the set of all possible market states as represented at different time instances through measured market indicators. Our ability to distinguish market states is constrained by available market indicators. In other words, two states with the same indicators are indistinguishable which leads to the partitioning of all states into identity classes or elementary sets. The size of the elementary sets reflects the degree of precision, or granularity of knowledge representation. The degree of the granularity can be controlled in the DATALOGIC/R system by setting the "roughness" parameter in the range <0,1>. The finest classification is achieved when roughness is set to 0 and the coarsest partitioning is produced when roughness is 1. With the low granularity level the system tends to discover relatively many rules with weak support (strength) but high estimated decision probability. Such rules are likely to be generally incorrect due to the small amount of supporting evidence. When the roughness is high, the system typically identifies much fewer stronger and simpler rules but the estimated decision probabilities are normally lower. These rules are likely to be correct or almost correct but their usefulness depends on the magnitude of the probability of the predicted outcome.

4.2. LOWER BETA APPROXIMATION OF A SET

The lower BETA-approximation of a set (a concept) X in a given approximation space is a union of all elementary sets whose degree of overlap (i.e. the size of the intersection) with the set X is less or equal to BETA. Since the rules are produced based on the identified lower approximation, the parameter BETA in practice represents the minimum admissible, or user acceptable rule probability. The user of the DATALOGIC/R system can control the level BETA in search for the best compromise rules i.e. ones with relatively high strength and probability.

5. STOCK AND ECONOMIC DATA

Data was accumulated on a monthly basis from 1980 to 1990 from Statistics Canada and The Investment Corporation and put in a Lotus spreadsheet format. In total 120 information vectors with 40 attribute values each were collected and used in the analysis. The appendix has examples of our raw and discretized data. Through the use of Lotus macros the original raw data were discretized by replacing the values of the market indicators recorded for a given month with range symbols representing percentage change of the current value relative to the value recorded in the previous month. The discretization procedure has been determined by stock market expert. According to the expert the most appropriate way of assigning range symbols is as follows:

Range Symbol	% Difference
2:	20% to 29%
1:	10% to 19%
0:	0% to 9%
-1:	-10% to 0%
-2:	-20% to -11%

Some data was already given in a monthly % change format which needed a finer granularity range variable index defined as follows:

Range Variable	% Difference
2:	2% to 2.9%
1:	1% to 1.9%
0:	0% to .9%
-1:	-1% to 0%
-2:	-2% to -1.1%

The above two range indexes were used with the scope of the range symbols changing from -40 to 40. For example, a 53% change according to the first index produced the symbol 5. The converted data was exported into an ascii file and then imported into DATALOGIC/R. Before the import into DATALOGIC/R each variable had to be defined with a name, variable type, and length. Also, the tuning and analysis parameters such the roughness or the minimum acceptable rule probability parameter BETA with various reporting options had to be adjusted to provide different views of the information and thus possibly generating new hidden knowledge or confirming current knowledge.

The data being used can be broken down into the following areas of interest:

- stock market data.
- economic indicators data.
- individual company data.

The stock market data include the following indexes which represent the closing quotations at month-end:

- TSE - Toronto Stock Exchange Index.
- DOW - Dow Jones Index.
- S&P - Standard and Poors Index.
- P/E - Price Earnings Ratio for the TSE.
- 7 of the 14 major industry indexes:
  - Oil and Gas Index.
  - Metals and Minerals Index.
  - Utilities Services Index.
  - Paper and Forest Index.
  - Mechanising Index.
  - Financial Services Index.
  - Precious Metals Index.

The economic indicators include the following indexes which are represented as monthly percentage changes in most cases:

- M1 - Money available for investment purposes.
- Gross domestic product(GDP) in current prices.
- GDP at constant prices.
- Non-farm GDP.
- Industrial production in constant prices.
- Wages and salaries per unit of output.
- Total labour income.
- Corporate profits before taxes.
- Labour force total.
- Labour force employed.
- Government expenditures on goods and services.
- Non-residential fixed investment.
- Manufactures' inventories.
- Housing starts.
- Passenger car sales.
- Merchandise exports.
- Merchandise imports.
- Government of Canada - Interest rates.
- Security yield for Treasury bills - 3 month.
- Security yield for Canada bonds over 10 years.
- US dollar in Canadian dollars, avg noon rate.
- Unemployment rate.
- Consumer Price Index - Inflation rate.

The individual company stock data includes the following companies which represent the closing quotations at month-end:

- BCE Inc.

-Northern Telecom Incorporated.
-Bank of Montreal.
-Loblaw Inc.
-Imperial Oil.

6. THE DISCOVERED RULES

The following are the rules discovered from the 80's stock and economic data. By setting the roughness parameter one is able to view information differently. By setting the roughness parameter low, strong generalized rules are generated. By setting the roughness parameter high (1), exact (i.e. with estimated probability equal to 1) but relatively weak rules are identified. In all generations the minimum acceptable rule probability parameter BETA was set at 0.55. In the presented rules the "p" indicator represents probability and the "c" indicator represents the number of supporting cases (rule strength). The following are the results:

Bank of Montreal

Generalized Rules

1. This stock rises 0% to 10% when: (p=0.78,c=59)
   -the financial index fluctuates by -10% to 10%.
2. This stock drops up to 10% when: (p=0.78,c=54)
   -the financial index drops from 10% to 20%.

Exact Rules

1. This stock rises 0% to 10% when: (c=17)
   -the TSE P/E fluctuates by -10% to 10% and
   -the Imperial Oil stock is $38-$61 and
   -the corporate profits fluctuates by -24% to 8% and
   -the financial index fluctuates by -10% to 10%.
2. This stock rises 0% to 10% when: (c=15)
   -the utility index doesn't drop more than 10% and
   -the merchandise index hovers between -10% and 10% and
   -the non-farm GDP changes from -5% to 3% and
   -the corporate profits change from -24% to 8%.

Bell Canada Enterprises

Generalized Rules

1. This stock rises 0% to 10% when: (p=0.89,c=68)
   -utility index drops more than 30% or
   -the utility index rises by more than 10%.
2. This stock drops up to 10% when: (p=0.88,c=51)
   -the utility index fluctuates from -30% to 10%.

Exact Rules

1. This stock rises 0% to 10% when: (c=22)
   -the utility index fluctuates from -10% to 10% and
   -the interest rates are at 10% to 15% and
   -the paper index fluctuates from -20% to 20% and
   -the GDP fluctuates 4% to 7%.
2. This stock drops up to 10% when: (c=17)

-the utility index fluctuates from -20% to 10% and
-the GDP rises from 8% to 10% and
-the interest rate is less than 10% or > 15%.

Imperial Oil

Generalized Rules

1. This stock rises 0% to 10% when: (p=0.74,c=50)
   -the oil index fluctuates from -10% to 10%.
2. This stock rises 10% to 20% when: (p=0.56,c=9)
   -the precious metals and oil index rise 10% or more.
3. This stock drops up to 10% when: (p=0.81,c=33)
   -the oil index fluctuates from -30% to 20% and
   -the TSE P/E fluctuates from -20% to 10%.

Exact Rules

1. This stock rises 0% to 10% when: (c=12)
   -inflation is 4% to 8% and
   -the DOW drops more than 10% or rises more than 10% and
   -the oil index drops up to 20%.
2. This stock drops up to 10% when: (c=11)
   -TSE drops more than 10% or rises more than 10% and
   -government expenditures is < 4% or > 13% and
   -interest rates are not at 10% or 11% and
   -the oil index fluctuates from -30% to 10%.

Loblaws

Generalized Rules

1. This stock rises 0% to 10% when: (p=0.83,c=28)
   -the finance index fluctuates from -10% to 10% and
   -the month is from November to April.
2. This stock rises 0% to 10% when: (p=0.82,c=17)
   -the month is from October to February and
   -the industrial production fluctuates from -2% to 4%.

Exact Rules

1. This stock rises 0% to 10% when: (c=12)
   -the finance index fluctuates from -10% to 10% and
   -the month is from October to June and
   -the industrial production is < 2% or > 4% and
   -the total labour does not change 1% or 2%.
2. This stock rises 0% to 10% when: (c=10)
   -the finance index fluctuates from -10% to 10% and
   -the DOW drops up to 10% and
   -the precious metals fluctuate from -20% to 10% and
   -the month is from January to April.

Northern Telecom

Generalized Rules

1. This stock rises 0% to 10% when: (p=0.68,c=34)
   -the S&P doesn't drop more than 10% and
   -the DOW fluctuates from -10% to 10% and
   -the car sales fluctuate from -24% to 8%.
2. This stock drops up to 10% when: (p=0.63,c=27)
   -car sales drop > 32% or rise > 8% and

-the paper index doesn’t fluctuate from 10% to 20%.

Exact Rules

1. This stock drops up to 10% when: (c=11)
-precious metals index fluctuates from -30% to 10% and
-paper index doesn’t fluctuate from 10% to 20% and
-car sales do not fluctuate from -31% to 9% and
-corporate profits fluctuate from -16% to 32%.
2. This stock drops up to 10% when: (c=8)
-S&P fluctuates from 20% to 10% and
-metals index fluctuates from -10% to 10% and
-corporate profits fluctuate from -16% to 32%.

7. EXPERT’S EVALUATION OF THE RULES

In this section we present comments of our expert, an experienced stock broker, Vice President and Director of First Marathon Securities Ltd., Mr. Donald Edwards regarding the quality of the data used and the utility and reasonableness of the rules discovered by the DATALOGIC/R system.

7.1. STOCK BROKER’S COMMENTS

> *" Before getting too involved in a discussion of the "Rules" that this program has generated, let me first comment on the data used and the variable range.*
>
> *There is no question about the quality of the data. The sources are irrefutable and easily verified. The only difficulties with the data are the quantity, more accurately the lack of quantity, and the fact that some of the data is only recorded on a quarterly basis. More data with greater frequency would certainly add to the reliability of the rules.*
>
> *The other area of concern would be the variable range chosen. Given the data available this chosen variable ranges are probably very good. With more information available the variable range will have to be narrowed.*
>
> *Given all of the above, this program was still able to develop some amazing rules and discover many logical relationships from the generalized rules. Generally speaking I am very impressed with the "Generalized Rules". These are almost all recognized rules or relationships in the investment industry. The "Exact Rules" are much more difficult to interpret.*
>
> *On a stock-by-stock basis:*
> ***Bank of Montreal*** *- The rules recognize the relationship of this stock to the financial index. It also appears that these rules have recognized the superior performance this stock can often have. This is often due to the fact that it*

has had a higher dividend.

Bell Canada Enterprises *- The generalized rules seem to have recognized the power of this stock as a member of the utilities index. In this index "Bell" has a heavy weighting and can sometimes drive the index. In addition, it is powerful enough that it can withstand downward pressure on the index without being pulled down as badly.*

Imperial Oil *- Of the "Generalized Rules" generated this stock has the most interesting. The first rule recognized Imperial Oil as a major component of the oil index. It also recognizes the greater degree of safety that a large integrated oil company was able to provide throughout the '80's. The second rule appears to recognize the relationship that existed between oil and gold. At one time, particularly after the ascent of OPEC, there was an adage "As goes oil, so goes gold". Oil being inflationary and gold being a hedge against inflation. What this rules shows to me is a confirmation from the gold index of a firming in the price of oil and, therefore, a firming in the price of oil stocks, i.e. Imperial Oil. The last rule is a bit more difficult to interpret. It seems to imply to me that there is less volatility in this stock in times of greater volatility in the markets. As an integrated oil company this is much as one would expect.*

Loblaws *- If there is any weakness in the "Generalized Rules" this is it. I must first point out that I am not as familiar with this company as I am with the others. Having said this, I had difficulty trying to interpret these rules. As a matter of fact, they look more like "Exact Rules" than "Generalized Rules".*

Northern Telecom *- The first rule seems to recognize a couple of facts. The first is that while Northern Telecom is a Canadian company, it is listed on the New York Stock Exchange and a large number of its shareholders are Americans and American institutions. We then see the stock rising along with a rise in the major American indices. It may seem odd to see a relationship to car sales. Car sales, however, are a barometer of consumer confidence. If people are prepared to purchase big ticket items, then they are probably confident about the future of their economy. This translates into a more optimistic attitude generally and buoyancy in the economy and in stock markets too.*

The rules have been able to identify what I believe to be some legitimate and logical relationships between stocks and the indexes to which they belong. In some cases, I would have to suggest that the program's interpretation of the rule was backwards. By that I mean that because of the *heavy weighting of the stock in its index it was actually the stock that moved the index and not the index that moved the stock as the rules seem to imply.*

While it is easy to knit pick and find fault, I can see that this program has generated some "real" rules. It will be very interesting to see what can be done with more data and tighter variable ranges. "

8. CONCLUSIONS AND FUTURE RESEARCH DIRECTIONS

As indicated by the expert, not all the rules discovered by the system were of high quality. Some of them were quite complex and difficult to interpret (e.g. all exact rules or imprecise rules for Loblaws). Of course, these rules reflected true properties of the available data, properties which turned out not to be true in general. However, one can notice that almost all the difficult to interpret rules are also relatively weak whereas the rules given high marks by the expert are strong (supported by many cases). This result seems to imply that strong rules are more likely to be correct. This raises an interesting question: How strong a rule should be in order to be reasonably certain that it is generally correct? This, and similar questions will the subject of further experimental research.

In addition to improving the rule selection process, the data being used can be improved immensely. The quantity of the data is definitely insufficient and this, most likely, underlies the lack of good generalized rules for Loblaws. This also indicates that the quality of the data needs to be improved by including more information. To improve the quality all quarterly data will be replaced by data with a better frequency. More stocks and economic data will be analyzed. The best leading, coincidental, and lagging economic indicators will be used. A better range variable indexing scheme will be devised while incorporating a moving average. Regarding additional information sources stock volume values need to be used along with the number of shares outstanding(the float). Along with the P/E we need the yield. It's crucial that we tap into information regarding insider information such as management ownership and trading of the specific company. It is critical to get all 14 TSE major industry stock index values. To sum up, the data can be enhanced significantly to hopefully give us more hidden knowledge regarding market predictability.

The concept of dividend earnings is an important part of determining the price of the stock as indicated by some of the stock market models which have been around for a long time. The concept of dividend earnings on a monthly basis for the individual company needs to be retrieved and included in with the analysis.

By aligning the stock's price with the stock and economic data from the previous month, rules can be generated indicating what happened the month before to cause the stock to go up or down or stay the same. This approach can be taken by going several months past and thus generating temporal rules while determining intermediate trends in the market. Temporal rules such as these can be generated:

> Imperial Oil Stock will rise by 15% next month when:
> -the WTI price stays constant for 3 months and
> -gold rises 10% in the last 6 months and
> -interest rates in the last 9 months drop by 3%.

This may lead to the possibility of predictability of intermediate and minor trends which Charles Dow himself has said are impossible to predict. The data could be viewed weekly, daily, quarterly, or yearly which could possibly lead to different rules for different cases. Major, intermediate, and minor trends could possibly have some form of reasoning to them if the proper data is analyzed in the proper fashion. Predictability to this level of detail would give the investor a remarkable advantage in stock market investing.

Another related research problem is the experimental validation of the rules. The correctness of the estimated rule probabilities should be tested on an independent set of test cases from outside the data table used to produce the rules. This kind of testing will be conducted when larger quantities of market data will become available.

We have only touched on the areas with which we are familiar. Further experimental research should unveil other information that could prove invaluable in analysis and prediction. The search continues.

ACKNOWLEDGMENTS

The research reported in this paper has been partially supported by an operating grant from Natural Sciences and Engineering Research Council of Canada. The permission to use the test version of DATALOGIC/R software developed by Reduct Systems Inc. is greatfully acknowledged. Many thanks to First Marathon Securities and the Investment Corporation for their resources and to Adam Szladow for his helpful comments and suggestions.

REFERENCES

1. Artificial Intelligence Applications on Wall Street. Proceedings of 1991 IEEE Conference.

2. Pawlak, Z. Rough Sets: Theoretical Aspects of Reasoning About Data. Kluver Academic Publishers, Dordrecht, The Netherlands, 1991.

3. Piatetsky-Shapiro, G. Discovery of Strong Rules in Databases. Proc. of IJCAI-89 Workshop on Knowledge Discovery in Databases, 264-274

4. Ziarko, W. Variable Precision Rough Sets Model. Journal of Computer and Systems Sciences (in print).

5. Ziarko, W. Analysis of Uncertain Information in the Framework of Variable Precision Rough Sets. Foundations of Computing and Decision Sciences (in print).

6. Szladow, A. DATALOGIC/R: Mining the Knowledge in Databases. PC AI, January 1993, 25-41.

7. Weiss, G. The New Rocket Science ; Welcome to the Future of Finance. Business Week. November 2, 1992. 131-140.

8. Holderness, M. Artificial Intelligence ; Can the Computer beat the Market?. Globe and Mail. January 2, 1993.

9. McGough, R. Fidelity's Bradford Lewis Takes Aim at Indexes With His 'Neural Network' Computer Program. The Wall Street Journal. Tuesday, October 27, 1992.