Available online at www.sciencedirect.com
SCIENCE DIRECT®
Decision Support Systems 37 (2004) 583–597
Decision Support Systems
www.elsevier.com/locate/dsw

Applying rough sets to market timing decisions

Lixiang Shen[a,*], Han Tong Loh[a,b]

[a] Design Technology Institute Ltd, Faculty of Engineering, BLK E4-01-07, National University of Singapore, 10 Kent Ridge Crescent, Singapore 119260, Singapore
[b] Department of Mechanical Engineering, National University of Singapore, Singapore 119260, Singapore

Available online 5 July 2003

Abstract

A lot of research has been done to predict economic development. The problem studied here is about the stock prediction for use of investors. More specifically, the stock market's movements will be analyzed and predicted. We wish to retrieve knowledge that could guide investors on when to buy and sell.

Through a detailed case study on trading S&P 500 index, rough sets is shown to be an applicable and effective tool to achieve this goal. Some problems concerning time series transformation, indicator selection, trading system building in real implementation are also discussed.

Keywords: Rough sets; Trading system; Indicator selection; Uncertainty

1. Introduction

With the emergence of data mining techniques, there are more and more applications appearing in the area of economic prediction, such as using genetic algorithms to choose optimal portfolio [4], selecting the real-world stocks using neural networks [25] and predicting the S&P 100 index using rough sets [45]. These new techniques become alternatives to conventional statistical methods and even perform better in some cases.

The problem studied here is about the stock prediction for investors' usage. More specifically, the stock market's movements will be analyzed and predicted. We wish to retrieve knowledge that would guide investors on when to buy and sell.

The method chosen to be our analyzing tool is rough sets by Pawlak [34]. Although it is a relatively new data mining tool, it has inspired many scholars to carry out research in adapting this theory to various domains [37,44,49,59,61–64].

Based on the notion of the existence of indiscernibility relation between objects, rough sets deal with the approximation of sets or concepts by means of binary relations. Compared to other methods used in financial area, such as discriminant analysis, univariate statistical method and linear probability model, this new method has the following advantages [8,16]:

- It is based on the original data only and does not need any external information;
- It is a tool suitable for analyzing not only quantitative attributes but also qualitative ones;

* Corresponding author. Tel.: +65-6874-8037; fax: +65-6773-6098.
E-mail address: dtislx@nus.edu.sg (L. Shen).

0167-9236/03/$ - see front matter © 2003 Elsevier B.V. All rights reserved.
doi:10.1016/S0167-9236(03)00089-7

- It discovers important facts hidden in data and expresses them in the natural language of decision rules;
- The set of derived decision rules gives a generalized description of the knowledge contained in database, eliminating any redundancy typical of the original data;
- The obtained decision rules are based on facts, because each decision rule is supported by a set of real examples;
- The results of rough sets are easy to understand, while the results from other methods (credit scoring, utility function and outranking relation) usually require an interpretation of the technical parameters, with which the user may not be familiar.

The applications of rough sets in economic and financial prediction can be divided into three main areas: business failure prediction, database marketing and financial investment. The corresponding references are given in Table 1, together with the rough set models applied. The applications are described here to show the diversity of the problems that rough sets can handle. A detailed review of applications of rough sets in financial domains can be found in Ref. [58].

Most rough sets applications are focused on classification problem. In those applications, the data sets applied to rough sets are temporal independent, which means the data set can be shuffled without affecting the final results. However, for time series, such as the medical history of the patients or historical data of a stock, the sequential information is important. How can rough sets be adapted to extract information from them? In this paper, the problems concerning time series forecasting using rough sets, such as converting time series to rough sets objects, discretization and rough sets model modification, are discussed. A trading system with good performance based on rough sets is presented.

Table 1
Major application areas of rough sets and their corresponding models

RS models	Business failure prediction	Database marketing	Financial investment
RSES			[3,5]
LERS		[35]	
DataLogic	[56]	[26,28]	[14,15,23, 38–40,45,63]
TRANCE		[11,20–22]	
ProbRough		[35,36]	
Dominance Relation	[16]		
RoughDas and ProFit	[8,51–54]		[55]
Hybrid Model	[2,17]		

2. Converting time series to rough sets objects

There is difference between the time series with and without real-time constraints. Without real-time constraints, the only thing that matters is the chronology of events—this is also the case when the time between each event is a constant. However, with real-time constraints, the interval between each event should also be taken into consideration. Obviously, this interval is various. When applying rough sets, these time series cannot be put into directly. There is a need to construct the suitable data from these time series which can be used by rough sets, we call them 'rough sets objects'. Baltzersen [3] presented two methods concerning this problem in his thesis. One is the 'mobile window method', which moves a window along the time series, the data points falling into the window are transferred to a rough sets object. This algorithm is a practical means of creating rough set objects from time series. It presents snapshots of the objects as the window moves. Baltzersen [3] recommended that the length of the window should be chosen based on the generalization of the Markov property, which means that the current value is only dependent on previous values by a limited number.

The other method is the 'columnizing method'. The time series are organized in columns, such that each row represents an object where each column is an economic indicator, and each row represents a different point in time. The detailed procedure consists of the following:

Step 1 Creating initial columns: The different indicators generated from time series are converted to adjacent columns.

Step 2 Synchronizing: The columns are shifted so that in each row, the same point of time is represented. For a stock price, moving average indicators may be calculated using different time intervals. They can be synchronized by shifting

one column against the other according to the time they are benchmarked.

Step 3 Dealing with null values: Appropriate methods for dealing with null values/mismatching frequencies are applied.

Step 4 Adding time attributes: Attributes identifying the time at which the events occur may be added. This attribute can be the decision attribute. For a stock price, trading signals can be added as time attribute.

Step 5 Creating derived series: Composing the decision table for rough sets application. (A decision table is one of formats that can be analyzed by rough sets. It is composed of columns labeled by attributes and rows by objects.)

It can be seen that this algorithm is more straightforward compared to the 'mobile window' method since the determination of 'markings' is difficult for us because of the expert's experience needed. In addition, the choice of the window length with Markov property is a time-consuming problem. So in the following experiment, the 'columnizing' method will be adopted.

3. Trading system based on rough sets

The most widely used methods in economic forecasting are fundamental and technical analyses. Fundamental analysis involves predicting stock price from other factors. For instance, predicting the stock price for an individual company may involve using figures from the company's annual report and indicators of the general economy. However, technical analysis only considers the actual history of trading and price in a security or index. The underlying theory is based on an assumption that the market price reflects all known information about the individual security.

Technical analysis is mainly concerned with market indicators. These technical indicators look at the trend of price indices and individual securities. They evaluate the current position of the security or index. The theory underlying these indicators is that once a trend is in motion, it will continue in that direction [1]. Technical analysis attempts to determine the strength of the trend and the direction of the trend. The technical analyst will use his analysis to stay away from a market or a security unless there is a good amount of protection in place.

Considering the data available for our experiment, our approach is a form of technical analysis. The trading system based on rough sets is illustrated in Fig. 1.

First of all, the historical data should be transformed to rough set objects which can be processed by RoughSOM algorithm. (RoughSOM is a new algorithm that combines Rough Set Theory and Self-Organizing Maps. The details will be given in Section

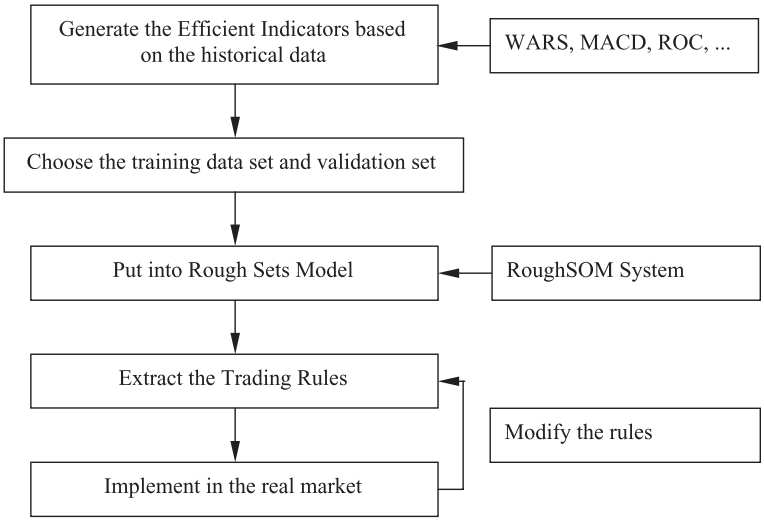

Fig. 1. The process of stock market data prediction and analysis.

3.2.) This has been done based on the 'columnizing method' introduced in Section 2. Altogether, seven indicators are chosen to compose the decision table that is used for reducts and rules generation. At this step, the decision table is processed using the Rough-SOM system, which will be introduced in the next section. This may be considered to be the "heart" of the entire process because at this step, there are many factors affecting the final results, such as rule selection and threshold values. These factors will be discussed in the experiment. The performance of the trading system is evaluated according to the net profit and Sharpe Ratio.

3.1. Indicators study

Before the application of rough sets, the historical price data needs to be converted to rough set objects. Following the 'columnizing method' introduced in previous section, the first step is to create the initial columns. Because the tools of the technical analysis are indicators and systems that are based on price charts, each column will correspond to an indicator that should be able to reflect the market change. In the following, the work on indicators' studies is presented.

One of the basic tenets put forth by Charles Dow in the Dow Theory [6] is that security prices do trend. Trends are often measured and identified by "trendlines" and they represent the consistent change in prices (i.e., a change in investor expectations). In Figs. 2 and 3, rising trend and falling trend are illustrated.

As shown in Fig. 2, a rising trend is defined by successively higher low-prices. A rising trend can be thought of as a rising support level—the "bulls" are in control and are pushing prices higher. Fig. 3 shows a falling trend. A falling trend is defined by successively lower high-prices. A falling trend can be thought of as a falling resistance level—the "bears" are in control and are pushing prices lower.

A principle of technical analysis is that once a trend has been formed, it will remain intact until broken [1]. The goal of technical analysis is to analyze the current trend using trendlines and then either invest with the current trend until the trendline is broken, or wait for the trendline to be broken and then invest with the new (opposite) trend. So for practical trading, to know the current market state, either in the rising trend or in the falling trend, is very important. Searching for indicators that can reflect the fluctuation of price in financial market is very important in our studies.

In our research, an indicator, called Weighted Accumulated Reconstruction Series (WARS), has been constructed and found to have interesting characteristics [43]. This indicator can reflect the trend of the changes in price. In addition, it makes use of more

Fig. 2. Rising trend.

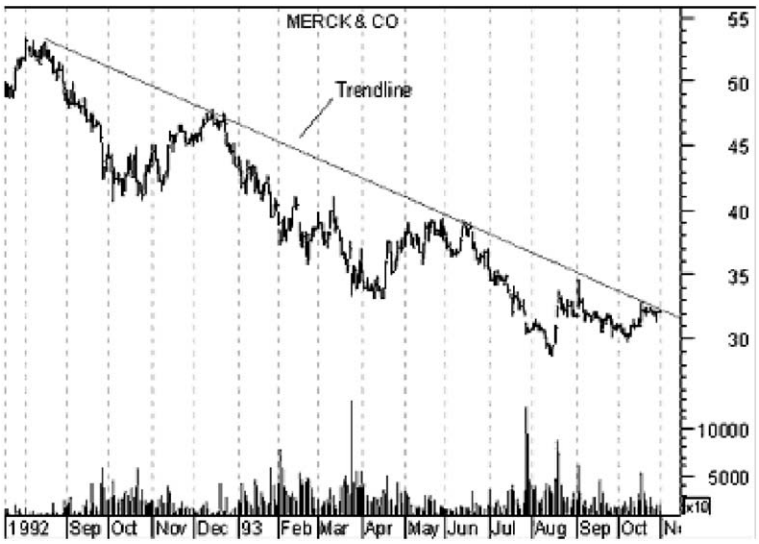

Fig. 3. Falling trend.

information contained in the data than the *moving average* and therefore may be able to better reflect the state of the price or index. The details about how to construct WARS and its performance in catching market trend can be found in Ref. [43].

Another six indicators are chosen from well-established ones in the technical analysis. They are Moving Average Convergence/Divergence (MACD), Price Rate of Change (ROC), Stochastic Oscillator, Relative Strength Index (RSI), Directional Indicator (DI) and Linear regression lines. Table 2 gives the definition of all six indicators.

Besides these six indicators, other indicators can be adopted as well. The proposed system here is a generic system that can use additional—or totally different—indicators.

3.2. RoughSOM system

Classifying a new object by matching its description to rules generated by the Rough Set Theory will lead to one of four situations that are commonly identified. They are as follows:

(i) The new object matches exactly one of deterministic sorting rules;
(ii) The new object matches exactly one of nondeterministic sorting rules;
(iii) The new object does not match any of the sorting rules;
(iv) The new object matches more than one rule.

In (i), the sorting suggestion is obvious. In (ii), however, the suggestion is not direct since the matched rule is ambiguous. In this case, the Decision Maker (DM), who applies the Rough Set Theory to solve his own problem, is informed of the number of sorting examples that support each possible category. This number is called the *strength*. If the *strength* of one category is greater than the *strength* of other categories occurring in the nondeterministic rule, one can conclude that according to this rule, the considered object most likely belongs to the strongest category.

Situation (iii) is more difficult to solve. In this case, one can help the DM by presenting him with a set of rules 'nearest' to the description of the new object. The notion of 'nearest' involves the use of the distance measure. Slowinski [50] has proposed a distance measure based on a *valued closeness relation R* having some good properties.

Situation (iv) may also be ambiguous if the matched rules (deterministic or not) lead to different classes. Here, the suggestion can be based either on the *strength* of possible classes, or on an analysis of the sorting examples, which support each possible class.

Table 2
Definitions of the six indicators

Indicators	Definition
MACD	$\sum_{i=1}^{n} \mathrm{EMA}_{20}(i) - \sum_{i=1}^{n} \mathrm{EMA}_{40}(i)$, where $\mathrm{EMA}_n(i) = a \times p(i) + (1-a) \times \mathrm{EMA}_n(i-1)$, $a = 1/n$
ROC	$(p(i) - p(i-n))/p(i) \times 100$
Stochastic Oscillator	$\sum_{i=1}^{n} (p(i) - h(i))/(h(i) - l(i))$
RSI	$\mathrm{RSI} = 100 - \frac{100}{1_n + \mathrm{RS}}$, where $\mathrm{RS} = \sum_{i=1}^{n} (p(i) - p(i-1))^{+} \Big/ \sum_{i=1}^{n} (p(i) - p(i-1))^{-}$
DI	$\mathrm{DI}^{+} = +\mathrm{DM/TR}$, $\mathrm{DI}^{-} = -\mathrm{DM/TR}$; DM—Directional Momentum, TR—True Range
Linear regression lines	$\left(n \times \sum_{i=1}^{n} i \times p(i) - \sum_{i=1}^{n} i \times \sum_{i=1}^{n} p(i)\right) \Big/ \left(n \times \sum_{i=1}^{n} i^2 - \left(\sum_{i=1}^{n} i\right)^2\right)$
WARS	[43]

$p(i)$ = close price; $h(i)$ = high price; $l(i)$ = low price; n = number of points.

In the latter case, the suggested class is that which is supported by a sorting problem closest to the new object based on the relation R.

In the solutions to situations (ii) and (iv) mentioned above, one situation is missing, that is, the situation whereby the *strengths* of both categories are the same in which case there is no clear indication on how to classify the new object. For situation (iii), from the view of *valued closeness relation* R, there are also some parameters such as the *relative importance* k_l and *veto threshold of criterion* c_l [50] which have to be predefined. This is not favorable to the users without sufficient experience. In order to enhance the sorting capability of rough sets and remove some ambiguities of the system, the current research is undertaken. For rough sets, although no exterior information is needed in generating the rules, the ''inner relationships'' among objects are ignored. ''Inner relationship'' refers to the category the object belongs to as determined by a cluster analysis. This ''inner relationship'' is helpful in distinguishing the ''strong'' objects from the ''weak'' objects. By ''strong'' object, one means that the inner category for the object as determined by cluster analysis is the same as the original decision class. Otherwise, the object is termed as a ''weak'' object. The information inherent in the objects helps to remove the uncertainty from the system and increase the sorting accuracy on the new objects. This is especially efficient for inconsistent systems. The Self-Organizing Maps (SOM) [19] is applied in this project as a cluster method.

Following this idea, a new method termed Rough-SOM, which combines Rough Set Theory and SOM, is proposed. The flow chart of this new algorithm is illustrated in Fig. 4.

This new algorithm is composed of three steps, namely, Decision Table Reconstruction, Discretization and Rough Sets Application. In the Decision Table Reconstruction step, the original training data set is categorized using the SOM. We constrain the output groups to be the number of decision values. Through the process on the training data set by SOM, the new decision value is obtained for every object. Now each object corresponds to two decision values. So far, there are fewer references on solving the multiple decision attribute problem. In our research, the original decision table together with a new decision attribute obtained by SOM is reconstructed according to the following rules:

- If the object's original decision value is the same as the one generated by SOM, then its decision value does not change.
- If the object's original decision value is different from the one generated by SOM, then this object is given a new decision value. For example, there are two original decision values, say 1 and 2, if original Decision = 1, SOM Decision = 2, then new Decision = 3; if original Decision = 2, SOM Decision = 1, then new Decision = 4.

The reason for this reconstruction is to ensure that the new decision table has the same *quality of approximation* as the original decision table.

The second step of this new algorithm is to discretize the continuous attributes in reconstructed decision table for the succeeding rough sets processing. At this step, there are a lot of well-established discretization methods [7,9,13,29–33,46,47,60]. The modified Chi2 algorithm is chosen as our discretization tool. This method, derived from ChiMerge algo-

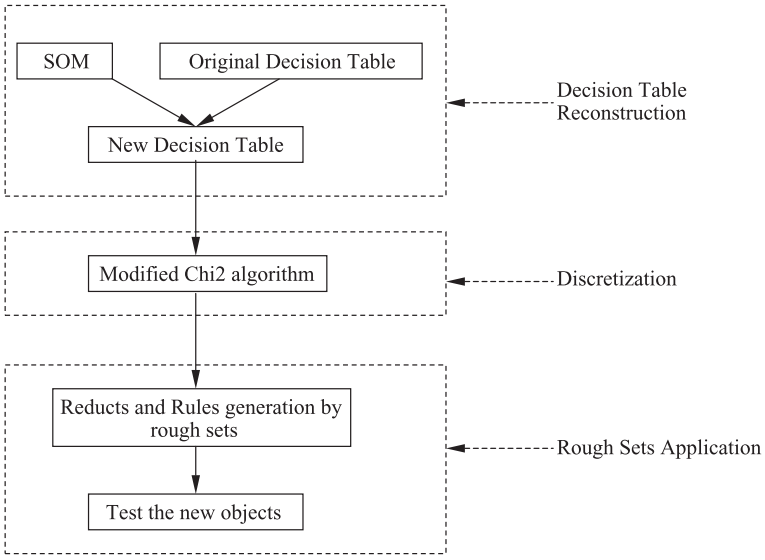

Fig. 4. The flow chart of RoughSOM algorithm.

rithm, was first proposed by Randy Kerber [18], as a statistically justified heuristic method for supervised discretization. Liu and Setiono [24] automated the discretization procedure by decreasing the significance level and keeping the inconsistency rate from exceeding a certain threshold. Tay and Shen [57] removed the inaccuracy existing in merging criterion and applied the *quality of approximation* coined from rough sets to avoid the user-defined threshold. With these two modifications, this algorithm has now become a fully automatic discretization method.

The last step is to apply rough sets to generate rules and check the predictive accuracy on the testing data sets. In this step, basic rough sets algorithm is used. In addition, the SOM, which is used to modify the support value of each object, is applied as well. As for which reduct to be selected to generate rules, Düntsch and Gediga [10] proposed a normalized entropy measure to do so. In our case, the following simple heuristic method is adopted. In case several reducts are generated by basic rough sets, the reduct with the minimal attribute number is selected. In addition, the reduct including the attribute with the highest *quality of approximation* is selected in the case there are several equal-length reducts. The reconstructed decision values in final rules will be transferred back to original decision values. However, at this time, the "strong" rules have been distinguished from the "weak" rule with their *strengths*.

Table 3 shows the predictive accuracy of 10 data sets chosen from UCI Machine Learning data base. The 10-fold cross-validation test method is applied to all data sets. The final predictive accuracy is taken as the average of the 10 predictive accuracy values with standard deviation associated. Through experiments carried out on these 10 data sets, it has shown this new method removes the uncertainty from the system and increases the predictive accuracy on the test data sets. However, for a data set with four or more classes,

Table 3
The predictive accuracy (%) of original rough sets and RoughSOM

Data set	Rough sets	
	Original	SOM
Australian	78.84 ± 6.42	84.20 ± 4.40
Bupa	53.53 ± 9.16	53.82 ± 14.74
Diabetes	70.26 ± 6.74	71.05 ± 6.54
Cleveland	55.86 ± 7.23	55.43 ± 6.98
Heart disease	79.26 ± 6.34	80.15 ± 4.43
Hsv-r	64.17 ± 15.74	64.17 ± 15.74
Iris	94.00 ± 7.34	96.00 ± 10.52
New thyroid	94.76 ± 9.38	95.90 ± 5.04
Vehicle	63.57 ± 4.53	63.13 ± 5.56
Wine	87.65 ± 18.23	88.82 ± 12.84
Average	74.19	75.27

such as the Cleveland Heart Disease, Hsv-r and Vehicle data sets, our predictive accuracy is not as good as the basic rough sets results. The reason is partly due to the uneven distribution of the data sets. This problem might be solved by selecting reduct using normalized entropy measure [10] and will be studied in the future.

To analyze the results obtained in Table 3, the Wilcoxon matched-pairs sign-rank test [27] is applied. The purpose of this nonparametric test is to determine if significant differences exist between the two populations. Paired observations from the two populations are the basis of the test, and the magnitudes of the differences are taken into consideration. This is a straightforward procedure to either accept or reject the null hypothesis that the two populations are identical population distributions. The result of applying the Wilcoxon matched-pairs sign-rank test is that the RoughSOM outperforms the original rough sets algorithm at 2.5% significance level for a one-tailed test. It can be seen that the new method is more accurate than the original rough sets algorithm.

4. Experiment

4.1. Data preparation

In Section 3.1, seven indicators have been presented to compose the condition attributes of the decision table. We also need to define the decision table that indicates future direction of the data set. The decision attribute is constructed as follows:

$$\text{Dec\_att} = \left\{ \sum_{i=1}^{20} (21 - i) \cdot \text{sign}[\text{close}(i) - \text{close}(0)] \right\} \Big/ \sum_{i=1}^{20} i \quad (1)$$

where close(0) is today's close price and close(i) is the ith close price in the future.

Eq. (1) specifies a range -1 to $+1$ for Dec_att. A value of $+1$ indicates that, every day, up to 20 days in the future, the market closed higher than today. Similarly, a -1 indicates that, every day, up to 20 days in the future, the market closed lower than today. Fig. 5 gives a snapshot of the S&P 500 index for the period covering from Jan. 1999 to Jul. 1999 and the fluctuation of the Dec_att.

In our analysis, the Dec_att takes three values, $+1$ corresponding to buy, 0 to hold and -1 to sell. Different unsupervised discretization methods are studied for assigning value of Dec_att according to the value of Sharpe Ratio [41] and trading numbers. Finally, the equal-frequency-interval method is chosen to implement the discretization. The Sharpe Ratio is defined as follows:

$$\text{Sharpe Ratio} = \frac{\text{mean of the returns}}{\text{standard deviation of the returns}} \quad (2)$$

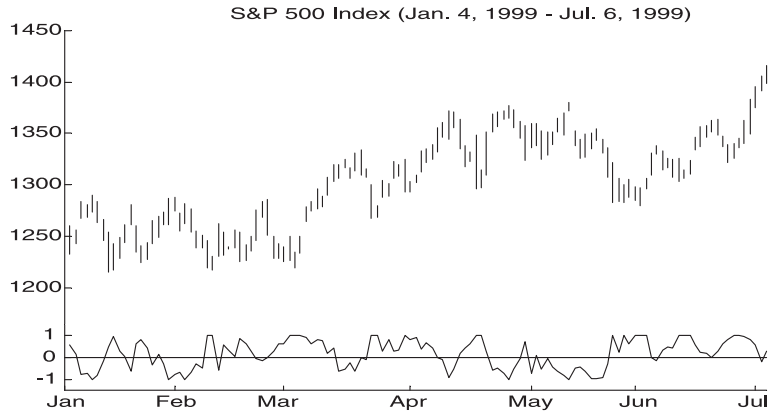

Fig. 5. S&P 500 index covering period Jan. through Jul. 1999. The Dec_att is shown below the S&P 500 index.

In our experiments, four future indices were chosen to be studied. They are S&P 500 index (Standard & Poor 500 stock index futures), MATIF-CAC index (French government stock index futures), EUREX-BOND (German 10-year government bond) and CBOT-US (United States 30-year government bond). The details on these four data sets are given in Table 4. There are two types of data used in our experiment. They are 15-min-bar data and daily-bar data, whose sampling frequency are 15 min and daily, respectively. The 15-min-bar data is used to calculate the condition attributes. Each indicator is calculated using the data points within 1 day so that each indicator reflects the fluctuation on the daily basis. The decision attribute is calculated using daily-bar data. In the end, the condition attributes and decision attribute keep the same time frequency. It is unnecessary to shift the data either in row or in column.

4.2. Rules extraction

After construction of decision table, these data sets are divided into a rule-extraction set which covers the period before 1999, and a validation set which covers the period after 1999. In both extraction and validation data sets, the rows of data corresponding to $1/6(\max\_wars - \min\_wars) < WARS < 5/6(\max\_wars - \min\_wars)$ are removed because our analysis is to find the distinguished rules to generate stronger trading signals. This threshold is set to limit the number of the training sets to be less than 1000 because the modified Chi2 algorithm is suitable for discretizing medium-size data set (<1000). So the threshold is various for different cases. Fig. 6 provides a snapshot of the S&P 500 index for the period covering from Jan. 1999 to Jul. 1999 and the fluctuation of the WARS.

Table 4
Data sets information

Data set	Starting date	Ending data	Number of points (before/after 1999)
S&P 500	Jan. 4, 1988	Jul. 26, 1999	2929 (2781/248)
MATIF-CAC	May 25, 1993	Jul. 6, 1999	1539 (1388/151)
EUREX-BOND	Jan. 2, 1991	Aug. 12, 1999	2155 (2002/153)
CBOT-US	Oct. 1, 1990	Jul. 6, 1999	2219 (2072/147)

The selected training data sets are discretized using the modified Chi2 algorithm. After discretization, the decision table is sent into the RoughSOM system to generate rules. Here the discernibility function is used to rule generation [48]. The format of rules is presented as follows.

Rule #1:	If	Price rate of change within a day < -0.632107	Then	Sell it in next morning
Rule #2:	If	$-0.508249 \le$ MACD within a day < -0.477388	Then	Sell it in next morning
⋮				
Rule #50:	If	$0.358209 \le$ directional index within a day < 0.577465 and $0.591435 \le$ Stochastic Oscillator within a day < 0.595588	Then	Buy it in next morning
Rule #51:	If	$0.154823 \le$ Linear regression lines within a day < 0.313858 and $0.591435 \le$ Stochastic Oscillator within a day < 0.595588	Then	Buy it in next morning
⋮				
Rule #300:	If	$1.441516 \le$ Price rate of change within a day < 1.740203 and $0.330097 \le$ Stochastic Oscillator within a day < 0.357194	Then	Buy it in next morning
Rule #301:	If	$0.726115 \le$ RSI indicator within a day < 0.887273 and $0.330097 \le$ Stochastic Oscillator within a day < 0.357194	Then	Buy it in next morning
⋮				

These obtained rules are used to build the trading system. Several runs using different settings, i.e., *strength threshold* and methods to solve unseen cases, for each of these are usually necessary in order to obtain a sense of the quality of the extracted rules. Here *strength threshold* is chosen according to rule number and the generalization of each rule. Several runs have been done to choose suitable threshold and the results are presented in Table 6. In case there is no rule matching the new objects, two methods are used. One is by calculating Euclidean distance, which measures

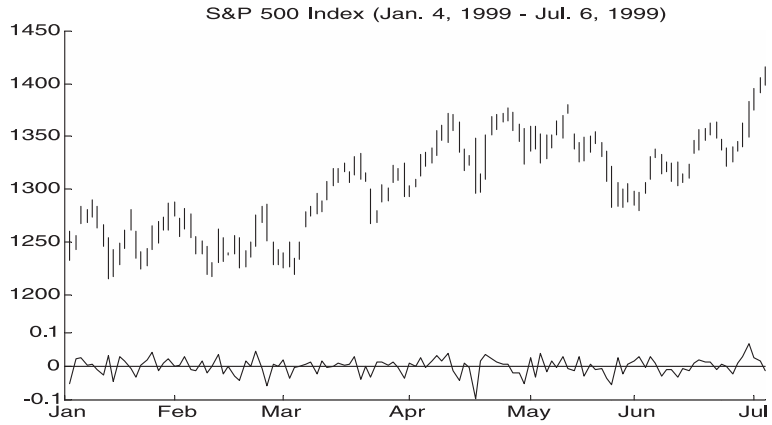

Fig. 6. S&P 500 index covering period Jan. through Jul. 1999. The WARS indicator is shown below the S&P 500 index.

the similarity between the training data and testing data; the category is assigned to this unseen object with the maximal frequency in the case where many equal distances are obtained. (Other distance measures, such as Hamming distance, lattice based distance, can also be applied here.) The other is to assign the unseen object a value 0 since in our special case of building trading system, 0 represents hold—meaning to take no action. In this way, the amount of trading will be reduced.

5. Results and discussion

The derived rules are encoded into a trading system. The performance of this trading system, for the period Jan. 1988 to Jul. 1999 of S&P 500 index is shown in Tables 5 and 6. The commissions and slippage have been ignored to simplify the study.

Table 5
Performance benchmark of buy–hold strategy and original decision attribute for period Jan. 4, 1999 to Jul. 26, 1999 of S&P 500 index

Method		Performance	
		1988–1998	1999
Buy–hold		1146.30	163.00
Dec_att	Net_profit	4624.50	539.60
	max_win	157.50	79.50
	max_loss	− 13.25	− 1.20
	Trading_number	323	15
	Winning_Trade	293	14
	Sharpe_Ratio	0.68	1.31

The results are compared to the buy–hold strategy. The reason to choose this simple strategy as benchmark is to beat Efficient Market Hypothesis [12]. The buy–hold strategy is defined as follows:

> *Buy–hold strategy:* A buy–hold strategy assumes that you buy on the first day loaded in the chart and hold the position until the end of period of study. The profit is calculated by using the price on the first day and the price on the last day.

From Tables 5 and 6, it can be seen that for the whole test period covering 1988 to 1999, the trading system built using rough sets always performs better than that of the buy–hold strategy because it generates more net profit. However, for the validation set, after 1999, it performs worse than the buy–hold strategy except that threshold equals to 1.0. The trading system loses money and gets a negative Sharpe Ratio.

Comparing the different performance corresponding to different *strength threshold*, it can be seen that with the increase of the threshold, the test accuracy rate over the whole period does not fluctuate too much, always around 0.58, which means about 58% objects can be classified correctly. However, this does not mean a high profit since the order of the buy, hold and sell can be reorganized. The best threshold is 1.0 in this case study with the highest profit and Sharpe Ratio. In addition, its performance in the validation set, after 1999, is also better than the others.

From Table 6, it also can be seen that the real trading number is much greater than that generated by

Table 6
Performance of the trading system for period Jan. 4 1999 to Jul. 26 1999 of S&P 500 index

Threshold	Test accuracy rate	Rule no.	Performance				
			Parameters	1988–1998		1999	
				Eu[a]	unEu[b]	Eu	unEu
0.5	0.57786	3307	Net_profit	1509.50	1403.20	33.10	22.10
			max_win	98.50	98.50	45.50	45.50
			max_loss	−31.00	−36.50	−31.00	−36.50
			Trading_number	777	745	42	40
			Winning_Trade	385	362	19	18
			Sharpe_Ratio	0.16	0.15	0.04	0.02
1.0	0.58542	1044	Net_profit	1692.60	1479.60	168.70	20.70
			max_win	94.00	94.00	57.50	44.70
			max_loss	−31.00	−37.70	−31.00	−36.50
			Trading_number	825	611	44	38
			Winning_Trade	418	310	23	18
			Sharpe_Ratio	0.17	0.18	0.16	0.02
1.5	0.58542	894	Net_profit	1563.60	1261.90	104.00	−79.70
			max_win	94.00	113.50	57.50	44.70
			max_loss	−32.00	−37.70	−32.00	−36.50
			Trading_number	849	557	46	36
			Winning_Trade	427	281	23	14
			Sharpe_Ratio	0.16	0.16	0.10	−0.10
2.0	0.58336	362	Net_profit	1418.80	1415.40	−40.60	−197.50
			max_win	94.00	214.50	57.50	44.70
			max_loss	−49.00	−40.50	−49.00	−40.50
			Trading_number	849	415	49	29
			Winning_Trade	435	213	23	9
			Sharpe_Ratio	0.15	0.19	−0.04	−0.30
2.5	0.58371	336	Net_profit	1390.60	1128.20	−69.00	−125.50
			max_win	94.00	214.50	57.50	44.70
			max_loss	−49.00	−40.50	−49.00	−40.50
			Trading_number	835	385	51	25
			Winning_Trade	427	189	24	8
			Sharpe_Ratio	0.15	0.16	−0.07	−0.22
3.0	0.58439	152	Net_profit	1323.90	1384.40	−115.40	−144.50
			max_win	94.00	206.30	57.50	75.00
			max_loss	−40.00	−88.30	−40.00	−59.10
			Trading_number	829	270	51	21
			Winning_Trade	432	149	22	6
			Sharpe_Ratio	0.14	0.23	−0.11	−0.22

[a] Eu—using Euclidean distance to determine the category of unseen objects.
[b] unEu—assign the unseen object a value 0.

Dec_att. In Fig. 12, the trading system covering the validation set at 3.0 threshold is plotted to be compared with the Dec_att.

Fig. 7 shows that some test trading signals are the same as the original signals, which is more favorable to the users. In addition, the test trading signals catch more turning points than the original signal does, which means this trading system is sensitive to the trend reverse and it catches more information from this reverse. Correspondingly, there are more trading signals generated from Jan. to Mar. 1999, which is more volatile period compared to Mar. to Jul. 1999. This is also favorable to the users. As it is known that volatile market creates a difficult trading condition for inexperienced users, our trading system provides them more chances to win. However, there are more trades than the original signals and some trading signals are generated after the turning points, which means it stays

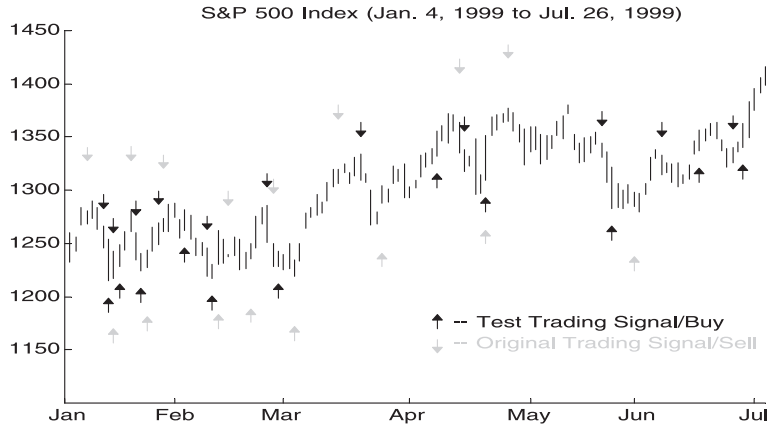

Fig. 7. Trading system by using threshold = 3.0 and by Dec_att covering period Jan. 4 to Jul. 26, 1999.

while the trends have changed from up to down or vice verse. From another point view, this may be caused by the noise in time series.

The performance of our trading system on three other stocks (MATIF-CAC index, EUREX-BOND and CBOT-US) can be found in Ref. [42]. The results support our previous analysis. For MATIF-CAC index, our system performs much better than buy–hold strategy. For EUREX-BUND and CBOT-US indices, compared to the negative profit by buy-hold strategy, our system cuts the loss and even makes profit by choosing correct threshold.

6. Conclusion

In this paper, the case study using rough sets to build a trading system in financial market is presented. Through the detailed analysis on S&P 500 index, it shows the procedure on how to apply rough sets to solve the temporal rule discovery problem. The problems concerning time series conversion to rough sets objects, indictor selection and removal of uncertainty from original rough sets are presented and discussed. The performance of the trading systems built by RoughSOM shows that this new knowledge discovery tool does find some inherent rules contained in the data. This can be seen from the Sharpe Ratio and net profit compared with the buy–hold strategy. The performance on three other future indices also supports our analysis. It is a promising alternative to the conventional methods. However, due to a lack of expertise, the generated rules cannot be evaluated and validated. Some trivial rules cannot be removed from the final trading rules. This is the possible reason why there are more trades activated compared with the original signals. We believe that if expert's experience is available, it will generate more promising results.

Although the trading system in this paper is based on rough sets, it can be applied to other algorithms as well, for instance, support vector machine (SVM) and artificial neural network (ANN) have made some progress in this area. Other discussed problems, such as time series transformation and indicators' selection, can also be applied to any similar or even totally different environment. We hope to see more data mining applications in this area.

Acknowledgements

We would like to thank the anonymous referees for their constructive remarks that helped to improve the clarity and the completeness of the paper.

References

[1] S.B. Achelis, Technical Analysis from A to Z: Covers Every Trading Tool—From the Absolute Breadth Index to the Zig Zag, Probus Publisher, Chicago, 1995.

[2] B.S. Ahn, S. Cho, C. Kim, The integrated methodology of rough set theory and artificial neural network for business failure prediction, Expert Systems with Applications 18 (2000) 65–74.

[3] J.K. Baltzersen, An attempt to predict stock market data: a rough sets approach, Diploma thesis, Knowledge Systems Group, Department of Computer Systems and Telematics, The Norwegian Institute of Technology, University of Trondheim, 1996.
[4] R.J. Bauer Jr., Genetic Algorithms and Investment Strategies, Wiley, New York, 1994.
[5] J.G. Bazan, A. Skowron, P. Synak, Market data analysis: a rough set approach, ICS Research Reports 6/94, Warsaw University of Technology, 1994.
[6] G.W. Bishop, H. Charles, Dow and the Dow Theory, Appleton-Century-Crofts, New York, 1960.
[7] M.R. Chmielewski, J.W. Grzymala-Busse, Global discretization of continuous attributes as preprocessing for machine learning, International Journal of Approximate Reasoning 15 (4) (1996) 319–331.
[8] A.I. Dimitras, R. Slowinski, R. Susmaga, C. Zopounidis, Business failure prediction using rough sets, European Journal of Operational Research 114 (1999) 263–280.
[9] J. Dougherty, R. Kohavi, M. Sahami, Supervised and unsupervised discretization of continuous features, Proceedings of the 12th International Conference of Machine Learning, Morgan Kaufmann, San Francisco, CA, 1995, pp. 194–202.
[10] I. Düntsch, G. Gediga, Uncertainty measures of rough set prediction, Artificial Intelligence 106 (1998) 109–137.
[11] A.E. Eiben, T.J. Euverman, W. Kowalczyk, F. Slisser, Modelling customer retention with statistical techniques, rough data models and genetic programming, in: A. Skowron, S.K. Pal (Eds.), Rough-Fuzzy Hybridization: A New Trend in Decision-Making, Springer, Berlin, 1998, pp. 330–348, Chapter 15.
[12] E.F. Fama, Random walks in stock market prices, Financial Analysts Journal, (downloadable from http://gsb.uchicago.edu/).
[13] U. Fayyad, K.B. Irani, Multi-interval discretization of continuous-valued attributes for classification learning, Proceedings of the 13th International Joint Conference on Artificial Intelligence, IJCA, Morgan Kaufmann, Los Angeles, CA, 1993, pp. 1022–1027.
[14] R. Golan, Stock market analysis utilizing rough set theory, PhD thesis, Department of Computer Science, University of Regina, Canada, 1995.
[15] R. Golan, D. Edwards, Temporal rules discovery using datalogic/R+ with stock market data, Proceedings of the International Workshop on Rough Sets and Knowledge Discovery (RSKD'93), Springer-Verlag, New York, 1993, pp. 74–81.
[16] S. Greco, B. Matarazzo, R. Slowinski, A new rough set approach to evaluation of bankruptcy risk, in: C. Zopounidis (Ed.), Operational Tools in the Management of Financial Risks, Kluwer Academic Publishing, Boston, 1998, pp. 121–136.
[17] R. Hashemi, L.A. Le Blanc, C.T. Rucks, A. Rajaratnam, A hybrid intelligent system for predicting bank holding structures, European Journal of Operational Research 109 (2) (1998) 390–402.
[18] R. Kerber, ChiMerge: discretization of numeric attributes, Proceedings of the 9th International Conference on Artificial Intelligence (AAAI-92), The AAAI Press, MIT, 1992, pp. 123–128.
[19] T. Kohonen, Self-organizing maps, Springer Series in Information Sciences, vol. 30, Springer, Berlin, 1995.
[20] W. Kowalczyk, Rough data modelling: a new technique for analyzing data, in: L. Polkowski, A. Skowron (Eds.), Rough Sets in Knowledge Discovery, vol. 1, Physica-Verlag, Heidelberg, 1998, pp. 400–421, Chapter 20.
[21] W. Kowalczyk, Z. Piasta, Rough set-inspired approach to knowledge discovery in business databases, Proceedings of the 2nd Pacific-Asia Conference (PAKDD-98), April 15–17, Springer, New York, 1998, pp. 186–197.
[22] W. Kowalczyk, F. Slisser, Modelling customer retention with rough data models, Proceedings of the 1st European Symposium (PKDD'97), Springer, New York, 1997, pp. 4–13.
[23] T.Y. Lin, A.J. Tremba, Attribute transformations on numeric databases and its applications to stock market and economic data, Proceedings of the 4th Pacific-Asia Conference (PAKD D2000), Springer, New York, 2000, pp. 181–192.
[24] H. Liu, R. Setiono, Feature selection via discretization of numeric attributes, IEEE Transactions on Knowledge and Data Engineering 9 (4) (1997) 642–645.
[25] G. Mani, K.K. Quah, S. Mahfoud, D. Barr, An analysis of neural-network forecasts from a large-scale, real-world stock selection system, Proceedings of the IEEE/IAFE 1995 Conference on Computational Intelligence for Financial Engineering (CIFER95), IEEE, New Jersey, 1995, pp. 72–78.
[26] D. Mills, Finding the likely buyer using rough sets technology, American Salesman 38 (8) (1993) 3–5.
[27] D.C. Montgomery, G.C. Runger, Applied Statistics and Probability for Engineers, 2nd ed., Wiley, 1999.
[28] A. Mrozek, K. Skabek, Rough sets in economic applications, in: L. Polkowski, A. Skowron (Eds.), Rough Sets in Knowledge Discovery, vol. 2, Physica-Verlag, Heidelberg, 1998, pp. 238–271, Chapter 13.
[29] H.S. Nguyen, S.H. Nguyen, Discretization methods in data mining, in: L. Polkowski, A. Skowron (Eds.), Rough Sets in Knowledge Discovery, vol. 1, Physica-Verlag, Heidelberg, 1998, pp. 451–482.
[30] H.S. Nguyen, S.H. Nguyen, Rough sets and association rule generation, Fundamenta Informaticae 40 (4) (1999) 310–318.
[31] H.S. Nguyen, A. Skowron, Quantization of real values attributes, rough set and boolean reasoning approaches, Proceedings of the Second Joint Conference on Information Sciences, Oct. 1995, Wrightsville Beach, NC, 1995, pp. 34–37.
[32] H.S. Nguyen, A. Skowron, Boolean reasoning for feature extraction problems, in: Z.W. Ras, A. Skowron (Eds.), ISMIS-97, Tenth International Symposium on Methodologies for Intelligent Systems, Foundations of Intelligent Systems, Charlotte, NC, USA, October 15–18, Lecture Notes in Artificial Intelligence, vol. 1325, Springer-Verlag, Berlin, 1997, pp. 117–126.
[33] H.S. Nguyen, S.H. Nguyen, A. Skowron, Searching for features defined by hyperplanes, in: Z.W. Ras, M. Michalewicz (Eds.), ISMIS-96, Ninth International Symposium on Methodologies for Intelligent Systems, Springer, New York, 1996, pp. 366–375.
[34] Z. Pawlak, Rough sets, International Journal of Computer and Information Sciences 11 (5) (1982) 341–356.

[35] D. Poel, Rough sets for database marketing, in: L. Polkowski, A. Skowron (Eds.), Rough Sets in Knowledge Discovery, vol. 2, Physica-Verlag, Heidelberg, 1998, pp. 324–335, Chapter 17.
[36] D. Poel, Z. Piasta, Purchase prediction in database marketing with the ProbRough System, Proceedings of the 1st International Conference (RSCTC'98), Springer, New York, 1998, pp. 593–600.
[37] L. Polkowski, A. Skowron (Eds.), Rough Sets in Knowledge Discovery, Vol. 1: Methodology and Applications. Vol. 2: Applications, Case Studies, and Software Systems, Physica-Verlag, Heidelberg, 1998.
[38] M. Ruggiero, How to build a system framework, Futures 23 (12) (1994) 50–56.
[39] M. Ruggiero, Rules are made to be traded, AI in Finance Fall (1994) 35–40.
[40] M. Ruggiero, Turning the key, Futures 23 (14) (1994) 38–40.
[41] W.F. Sharpe, The Sharpe Ratio, Journal of Portfolio Management Fall (1994) 49–58.
[42] L. Shen, Data mining techniques based on rough sets theory, PhD thesis, Department of Mechanical Engineering, National University of Singapore, 2003, can be downloadable from http://www.geocities.com/roughset/resume.html.
[43] L. Shen, E.H. Tay, Classifying market states with WARS, Proceedings of the 2nd International Conference on Data Mining, Financial Engineering, and Intelligent Agents (IDEAL 2000), Springer, New York, 2000, pp. 280–285.
[44] L. Shen, E.H. Tay, L. Qu, Y. Shen, Fault diagnosis using Rough Sets Theory, Computers in Industry 43 (1) (2000) 61–72.
[45] C. Skalko, Rough sets help time the OEX, Journal of Computational Intelligence in Finance 4 (6) (1996) 20–27.
[46] A. Skowron, Rough sets in KDD, in: Z. Shi, B. Faltings, M. Musen (Eds.), Proceedings of Conference on Intelligent Information Processing (IIP2000), Publishing House of Electronic Industry, Beijing, 2000, pp. 1–17.
[47] A. Skowron, Rough sets and Boolean reasoning, in: W. Pedrycz (Ed.), Granular Computing, Studies in Fuzziness and Soft Computing, vol. 70, Physica-Verlag, Heidelberg, 2001, pp. 95–124.
[48] A. Skowron, C. Rauszer, The discernibility matrices and functions in information systems, in: R. Slowinski (Ed.), Intelligent Decision Support—Handbook of Applications and Advances of Rough Sets Theory, Kluwer Academic Publishing, Boston, 1992, pp. 331–362.
[49] R. Slowinski (Ed.), Intelligent Decision Support—Handbook of Applications and Advances of the Rough Sets Theory, Kluwer Academic Publishing, Boston, 1992.
[50] R. Slowinski, Rough set learning of preferential attitude in multi-criteria decision making, in: J. Komorowski, Z.W. Ras (Eds.), Methodologies for Intelligent System, Springer-Verlag, Berlin, 1993, pp. 642–651.
[51] R. Slowinski, C. Zopoundis, Rough-set sorting of firms according to bankruptcy risk, in: M. Paruccini (Ed.), Applying Multiple Criteria Aid for Decision to Environmental Management, Kluwer Academic Publishing, Boston, 1994, pp. 339–357.
[52] R. Slowinski, C. Zopounidis, Application of the rough set approach to evaluation of bankruptcy risk, Intelligent Systems in Accounting, Finance and Management 4 (1995) 27–41.
[53] R. Slowinski, C. Zopounidis, A.I. Dimitras, Prediction of company acquisition in Greece by means of the Rough Set approach, European Journal of Operational Research 100 (1997) 1–15.
[54] R. Slowinski, C. Zopounidis, A.I. Dimitras, R. Susmaga, Rough set predictor of business failure, in: R.A. Ribeiro, H.J. Zimmermann, R.R. Yager, J. Kacprzyk (Eds.), Soft Computing in Financial Engineering, Physica-Verlag, New York, 1999, pp. 402–424.
[55] R. Susmaga, W. Michalowski, R. Slowinski, Identifying regularities in stock portfolio tilting, Interim Report IR-97-66, International Institute for Applied Systems Analysis, 1997.
[56] A. Szladow, D. Mills, Tapping financial databases, Business Credit 95 (7) (1993) 8.
[57] E.H. Tay, L. Shen, A modified Chi2 algorithm for discretization, IEEE Transactions on Knowledge and Data Engineering 14 (3) (2002) 666–670.
[58] E.H. Tay, L. Shen, Economic and financial prediction using rough sets model, European Journal of Operational Research 141 (2002) 643–661.
[59] S. Tsumoto, H. Tanaka, Automated discovery of medical expert system rules from clinical databases based on Rough Sets, Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD'96), The AAAI Press, Menlo Park, CA, 1996, pp. 63–69.
[60] H. Wang, I. Düntsch, G. Gediga, Classificatory filtering in decision systems, International Journal of Approximate Reasoning 23 (2000) 111–136.
[61] E. Weiss, Rough sets, rough neurons, induction and data mining, Neurovest Journal 4 (6) (1996) 28.
[62] E. Weiss, Rough sets, rough neurons, induction and data mining #2, Journal of Computational Intelligence in Finance 5 (3) (1997) 10–11.
[63] W. Ziarko (Ed.), Rough Sets, Fuzzy Sets and Knowledge Discovery, Proceedings of International Workshop on Rough Sets and Knowledge Discovery (RSKD'93), Springer-Verlag, New York, 1994.
[64] W. Ziarko, R. Golan, D. Edwards, An application of Datalogic/R knowledge discovery tool to identify strong predictive rules in stock market data, Proceedings of AAAI Workshop on Knowledge Discovery in Databases, Washington, DC, USA, The AAAI Press, Menlo Park, CA, 1993, pp. 93–101.

Lixiang Shen works at Design Technology Institute Limited, National University of Singapore. He received his PhD degree from National University of Singapore in 2002. Currently, his research focuses on applying data mining techniques to automate product development process. He is the main author and co-author of several journal papers. He is also the author of the book *Ordinary Shares, Exotic Methods*, which presents useful applications of data mining techniques in financial market. He is the main team member of KDDCup 2002 (Task 1) Honourable Mention Winners.

Han Tong Loh is the Deputy Director (Education) of the Design Technology Institute Limited. He is an Associate Professor in the Department of Mechanical Engineering at the National University of Singapore (NUS). He is also a Fellow of the Singapore-MIT Alliance, which is an innovative engineering education and research collaboration between MIT, NUS and NTU to promote global education and research in engineering. A/Prof. Loh's research interests are in the areas of data mining, rapid prototyping, robust design and computer aided design. He has published extensively in these areas.