Time-Series Data Analysis with Rough Sets

Joseph Herbert **JingTao Yao**

Department of Computer Science
University of Regina, Saskatchewan
E-mail: {herbertj,jtyao}@cs.uregina.ca

Abstract

The analysis of time-series data is important in many areas. Various tools are used for financial time-series data and there is no consensus for the best models. Rough sets is a new mathematical theory for dealing with vagueness and uncertainty. We apply rough set theory in the analysis of New Zealand stock exchanges. A general model for time-series data analysis is presented. The experimental results show that forecasting of the future stock movement, with reasonable accuracy, could be achieved with rough rules obtained from training data.

1. Introduction

Time-series data analysis is an important research domain of natural sciences, economics, and financial trading. It is crucial to certain applications of data mining, machine learning, and others in computer science. This type of analysis plays an important role in forecasting future circumstances. Forecasting is essentially taking historical time-series data, analyzing patterns and relationships between it, and producing a system or acquiring results that facilitate the prediction of future events of like situations [12, 14]. Many tools are available to assist in the analysis of this data but no consensus on which tools are the best.

Time-series data often posses content that is conflicting and redundant. The data may be imprecise, therefore a precise understanding of information cannot be derived from the data. If a precise understanding is not present, the use of hard computing techniques for the data analysis is difficult. Using soft computing techniques, or methods that do not ignore uncertain aspects of information, may help solve this problem. Rough sets may offer a powerful toolset for confronting this situation by being able to deal with such contradictory and superfluous records[8].

Using time-series data collected from the New Zealand stock exchange, a rough set model for the analysis of uncertain and redundant data will be used. The model that is used in this study follows the Knowledge Discovery in Database process given in [1, 3] closely.

2. Rough Sets and Applications

Rough set theory is a way of representing and reasoning imprecision and uncertain information in data [7, 16]. Based on the concept of indiscernibility, meaning that of the inability to distinguish between objects, rough set theory deals with the approximation of sets constructed from empirical data. This is most helpful when trying to discover decision rules, important features, and minimization of conditional attributes. There are four important concepts to discuss when talking about rough set theory: information systems, indiscernibility, reduction of attributes, and dependency.

Decision Tables (DT) are of the form $T = (U, C, D)$, where U is a universe made up of a non-empty finite set of objects, C is a set of conditional attributes, and D is a decision attribute. The union $A = C \cup D$ creates an attribute set consisting of both condition and decision attributes. Table 1 is an example of a DT.

Table 1. A Decision Table (DT)

Object	Index1	Index2	Index3	Decision
o_1	56.21	41.22	87.10	1
o_2	52.11	45.55	87.55	0
o_3	55.21	42.33	80.42	0
o_4	55.21	42.33	80.42	1

Discerning objects from each other is a major goal in rough set theory. For example, how are objects o_3 and o_4 in Table 1 discerned from each other? The values for conditional attributes are exactly the same, however, their decision based on those attributes are quite different. This is conveyed by saying that for any subset $B \subseteq A$,

$$IND_T(B) = \{(o, o') \in U^2 | \forall a \in B, a(o) = a(o')\}$$

where $IND_{IS}(B)$ is called the B-indiscernibility relation. Therefore, if $(o, o') \in IND_T(B)$, object o and o' are indiscernible from each other by attribute set B.

One of the important aspects in the analysis of decision tables extracted from data is the elimination of redundant attributes and identification of the most important attributes. Redundant attributes are any attributes that could be eliminated without affecting the degree of dependency

between remaining attributes and the decision. The degree of dependency is a measure used to convey the ability to discern objects from each other. The minimum subset of attributes preserving the dependency degree is termed *reduct* [7, 8]. A reduct of knowledge is its essential part, which suffices to define all basic concepts occurring in the considered knowledge.

Formally speaking, a subset $B \subseteq C$ is a reduct of C if B is independent and $IND(B) = IND(C)$. An attribute is considered independent if no loss of discerning power is made from the removal of this attribute. Obviously C may have many reducts. Obtaining reducts from time-series data allows knowledge to be of minimum length. For example, a DT with fifty conditional attributes and one decision attribute. After computing reducts, the decision D is be sufficiently discerned using only three of those attributes. Thus, performing reducts on the data can significantly reduce the amount of information required in order to discern between decisions.

Rough sets have been used in numerous applications, including feature selection [4, 17], fault diagnosis and other neural network applications [5, 11], forecasting [10], and other areas. The analysis of stock market time-series data using rough sets has been done with moderate success [9, 12, 13, 15]. The theory's versatility at classifying data is readily shown in numerous research endeavors.

3. The Rough Set Data Analysis Model

The model used in this study consists of three stages: data preparation, rough set analysis, and validation. Data preparation includes tasks such as data cleaning, completeness, correctness, attribute creation, attribute selection, and discretization. The rough set analysis generates preliminary knowledge, such as decision rules. This step is the first to directly create usable knowledge. The validation step confirms and filters knowledge with the validation data set.

The time-series data used was the stock market data from the New Zealand Exchange Limited (NZX). The NZX data begins July 31, 1991 and ends April 27, 2000. Data representing the closing price, opening price, highest price reached during the day, and the lowest price reached during the day.

3.1. Data Preparation

In order to successfully analyze data with rough sets, a decision table must be created. This is done with data preparation. The data preparation task includes data conversion, data cleansing, data completion checks, conditional attribute creation, decision attribute generation, discretization of attributes, and data splitting into analysis and validation subsets. Data conversion must be performed on the initial data into a form in which specific rough set tools can be applied.

Computational finance techniques offered methods to discover new conditional attributes that conveyed statistical properties of the data [14]. The following statistics were used: Moving Average Convergence/Divergence (MACD), Moving Average over a 5-day period (MA5), Moving Average over a 12-day period (MA12), Price Rate of Change (Price ROC), and Wilder Relative Strength Index (Wilder RSI). The decision attribute is the associated closing price of the next day, with values of either *-1*, *0*, or *1*. All attributes were discretized using an equal frequency algorithm. Data splitting created two subsets of size 1665 objects for the data analysis set and 555 objects for the validation set using a random seed.

3.2. Rough Set Analysis

Rough set analysis of data is used to obtain preliminary information. Methods such as approximation, reduct and core generation, classification, and rule generation can be performed on data to gain knowledge.

The rough set analysis of the data involved acquiring reducts (a minimum subset of attributes where the indiscernibility relation was unchanged), calculated from a subset of the original data. Rules were generated from the acquired reducts. The Rosetta Rough Set Toolkit [6] was used to perform reducts and generate decision rules. A filter was used to gather only those rules that met a minimum support threshold. Filtering a set of ninety-six original rules produced a subset of ten candidate decision rules.

Table 2. Reducts Generated

#	Reduct
1	{MA5, Price ROC, Wilder RSI}
2	{MACD, MA5, MA12, PriceROC}
⋮	⋮
17	{MA5, MA12, Price ROC}
18	{MA12, Price ROC, Wilder RSI}

The reducts that were generated from the data analysis are shown in Table 2. From the NZX data, a set of 18 reducts was obtained. Some reducts were unable to generate rules with sufficient support of the data and were filtered out.

The following is an example of a rule obtained from a reduct:

Rule 1:
IF MA5 = 1 and Price ROC = [*, -1.82045) and
Wilder RSI = [*, 37.17850)
THEN Decision3 = 1 or 0

Support (LHS) = [334 obj]
Coverage (LHS) = [0.200601]
Accuracy (RHS) = [0.931138]

The rule has three conditional attributes corresponding to the IF part. The rule has a decision of 1 (significant gain) *or* 0 (marginal change). From this rule, the conditional attributes have a support of 334 objects from a total of 1665 objects. Of those 334 objects, 93% have a decision value

of 1 or 0. Only rules with relatively high support and high accuracy are considered.

3.3. Validation

Once preliminary results have been obtained, validation procedures ensure that the knowledge is correct. If the preliminary knowledge is in the form of decision rules, these rules can be fired against the validation data set to confirm support, accuracy, and confidence measures. Rule filtering can be performed so that a subset of knowledge tailored to specific thresholds of the researcher is obtained. Comparison between measures obtained by firing the rules against the analysis and validation data subsets is needed to make certain that the knowledge is a correct depiction of the original data. Measures that are not similar with respect to the data set can be considered as being inferred from outlier or abnormal data.

4. Result Analysis

The preprocessed data was split into analysis (1665 objects) and validation (555 objects) sets. The analysis set was utilized by rough sets, which acquired reducts and decision rules. Each decision rule had measurements of support, accuracy, and coverage. These rules were the primary output of the rough set analysis process. The decision rules were then validated by firing each individual rule in conjunction with the validation data set. The new support, accuracy, and coverage measures were observed for the validation data set. The goal is to find rules that are accurate representations of the data. Therefore, rules that have similar measures in both the analysis and validation sets should be considered as stable and accurate.

It is interesting to see the result of combining two decision classes. Two possible combinations can be performed: the combination of a *decrease* and *neutral*, and the combination of an *increase* and *neutral*. The combination of *decrease* and *increase* potentially offers no value. However, the combination of *neutral* and either *increase* or *decrease* can tell us whether this rule can offer a large gain or loss with the inclusion of a marginal change. Rule 1 is an example of a rule with measurements obtained through both the analysis and validation sets.

This transformation results in the disjunction of two decision classes instead of one. It is interesting to see that the rule above can forecast either a large gain or a minimal loss or gain. This could be a powerful tool if the user so desired.

Each rule is applied to the validation data set corresponding to that of the analysis set. Information is collected regarding the support and accuracy of that rule in the new data.

The results shown in Table 3 indicate that for rule 1, the accuracy and coverage measures are extremely close according to the percentage of data used. This translates into a rule that is very stable. Rule 2, however, shows an accuracy measure of almost twenty percentage points higher using the validation set. This may need to be concluded as an unstable rule, since a margin of error of +/- 20% is considerably high.

Table 3. Reducts Generated

#	Data Analysis Set			Validation Set		
	Sup.	Acc.	Cov.	Sup.	Acc.	Cov.
1	334	0.646	0.201	114	0.649	0.205
2	97	0.659	0.058	34	0.823	0.061
3	206	0.577	0.123	58	0.603	0.104
4	48	0.688	0.029	18	0.611	0.032
5	112	0.5	0.067	35	0.714	0.063
6	327	0.605	0.196	122	0.581	0.219
7	112	0.634	0.067	33	0.606	0.060
8	252	0.615	0.151	78	0.564	0.141
9	111	0.596	0.067	34	0.618	0.061
10	86	0.488	0.052	25	0.520	0.045

The ten rules that were acquired through the process cover a total 1449 of 2220 objects in the data, 1090 and 359 objects in the analysis set and validation set respectively. Table 4 shows some statistical results.

Table 4. Statistical Results

	Analysis Set	Validation Set
Total objects	1665	555
Objects covered	1090	359
Min. support	48	18
Max. support	334	122
Average support	168.5	55.1
Min. accuracy	0.4884	0.5200
Max. accuracy	0.6875	0.8235
Average accuracy	0.6000	0.6291

Approximately 65.5% of distinct objects are covered by the ten rules in the analysis set. The same rules, when fired against the validation set, cover 64.7% of distinct objects in the validation set. Through the comparison of the accuracy and support measures given in Table 3, the amount of change in accuracy and support between the analysis set and validation set for each rule is shown in Table 5.

A large jump in accuracy from the analysis set to the validation set in rules 2 and 5 is observed. Although the support values for these two rules do not change significantly between data sets, this may tell us that these two rules are unstable since there is such a dramatic difference in accuracy. Rule 3 saw a significant decrease in support from the analysis set to the validation set. Whereas rule 3 covered 12.37% of objects in the analysis set, the same rule only covered 10.45% of objects in the validation set. Rule 1 is by far the strongest, most stable rule - showing negligible change in accuracy and support between data sets.

In order to determine which rules are the strongest, one must compare the rules and how they react with data. To analyze the rules that were have obtained, a ranking system could be used. Individual rules are ranked according to their strength (support and accuracy) and stability (change in support and change in accuracy). Taking into account all different types of rankings that were used, an overall rank can be determined. The results of this process can be seen in Table 6 where the header R1, R2, R3 and R4 are rank-

Table 5. Statistical Results

Rule	change in support (%)	change in accuracy (%)
1	0.480	0.242
2	0.300	16.374
3	-1.922	2.578
4	0.360	-7.639
5	-0.420	21.429
6	0.541	-2.354
7	-0.781	-2.787
8	-1.081	-5.098
9	-0.541	3.206
10	-0.661	3.163

Table 6. Ranking of Rules - Lower is Better

Rule	R1	R2	R3	R4	Final Rank
1	1	3	4	1	1
2	8	1	1	9	3
3	4	9	10	3	8
4	10	2	2	8	4
5	5	6	3	10	7
6	2	7	5	2	2
7	7	4	8	4	6
8	3	8	9	7	9
9	6	5	5	6	4
10	9	10	7	5	10

ings according to total support, total accuracy, change in support and change in accuracy respectively. The final rank is determined by the previous four rankings for each rule.

According to the rankings shown in Table 6, rule 1 is suggested as the best with low rankings for measures in support, accuracy, change in support, and change in accuracy. Rule 10 is considered the worst of the ten rules based on the high rankings of total support and total accuracy. Rules 4 and 9 both have a rank of four. This is due to the fact that rule 4 has high rankings for total support and change in accuracy but low rankings for accuracy and change in support. Rule 9 has average rankings for each measure.

A user may give different weights to different rankings. For example, if a user wishes to rank rules according to stability, he or she may incorporate a higher level of attentiveness to change in support and change in accuracy, whereas those measure contribute more to the final rankings of rules that those of total support and total accuracy. Equal weight was given to each measure.

5. Conclusion

The process of time-series data analysis could include the use of rough set tools for knowledge discovery. Using time-series data from the New Zealand stock exchange, rough set analysis acquired reducts used to create rough rules. The rules, after being tested for accuracy, are used as a forecasting tool to predict future configurations of data. Rough sets succeed in this task since they are able to describe uncertain data from information derived from precise, certain data. The data should be prepared so that it contains attributes that have minimal loss of information when they are discretized.

The results obtained seem to be quite stable, with some rules performing beyond expectations, especially those rules that were obtain by merging two decision classes together. This may indicate that rough sets can be a powerful tool for forecasting time-series data that is comprised of uncertain and redundant values. Acquiring reducts is beneficial to time-series data analysis, as it removes information that does not need to be present in order to discern objects or provide information about relationships between data.

References

[1] R. J. Brachman and T. Anand. The process of knowledge discovery in databases: A human-centered approach. In *Advances in knowledge discovery and data mining*, pages 37–58. American Association for Artificial Intelligence, 1996.

[2] B. S. Chlebus and S. H. Nguyen. On finding optimal discretizations for two attributes. *Lecture Notes in Computer Science*, 1424:537–544, 1998.

[3] U. M. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. From data mining to knowledge discovery: an overview. In *Advances in knowledge discovery and data mining*, pages 1–34. American Association for Artificial Intelligence, 1996.

[4] L. I. Kuncheva. Fuzzy rough sets: application to feature selection. *Fuzzy Sets Systems*, 51(2):147–153, 1992.

[5] H. Liu, H. Tuo, and Y. Liu. Rough neural network of variable precision. *Neural Process. Lett.*, 19(1):73–87, 2004.

[6] A. Ohrn. *ROSETTA Technical Reference Manual*, 2001.

[7] Z. Pawlak. *Rough Sets-theoretical Aspects of Reasoning about Data*. Kluwer Academic, 1991.

[8] Z. Pawlak. Vagueness - a rough set view. In *Structures in Logic and Computer Science, A Selection of Essays in Honor of Andrzej Ehrenfeucht*, pages 106–117, 1997.

[9] L. Shen and H. T. Loh. Applying rough sets to market timing decisions. *Decision Support Systems*, 37(4):583–597, 2004.

[10] T. G. Smolinski, D. L. Chenoweth, and J. M. Zurada. Application of rough sets and neural networks to forecasting university facility and administrative cost recovery. In *Artificial Intelligence and Soft Computing*, pages 538–543, 2004.

[11] W. Su, Y. Su, H. Zhao, and X. dan Zhang. Integration of rough set and neural network for application of generator fault diagnosis. In *Rough Sets and Current Trends in Computing*, pages 549–553, 2004.

[12] R. Susmaga, W. Michalowski, and R. Slowinski. Identifying regularities in stock portfolio tilting. Technical report, International Institute for Applied Systems Analysis, 1997.

[13] F. E. Tay and L. Shen. Economic and financial prediction using rough sets model. *European Journal of Operation Research*, 141:641–659, 2002.

[14] S. Taylor. *Modelling Financial Time Series*. John Wiley & Sons Ltd, 1986.

[15] Y. F. Wang. Mining stock price using fuzzy rough set system. *Expert Systems with Applications*, 24:13–23, 2003.

[16] Y. Yao, S. Wong, and T. Lin. A review of rough set models. In *Rough Sets and Data Mining: Analysis for Imprecise Data*, 1996.

[17] M. Zhang and J. T. Yao. A rough sets approach to feature selection. In *Proceedings of the 23rd International Conference of NAFIPS*, pages 434–439, 2004.