Research Journal of Mathematical and Statistical Sciences ISSN 2320–6047
Vol. **3(9),** 7-14, September **(2015)** *Res. J. Mathematical and Statistical Sci.*

Rule Based Mining of Nifty Fifty Stock Market Data Prediction Based on Rough Set Theory

T. Chitrakalarani and S. Indrakala
Kunthavai Naacchiyar Government Arts College for Women (Autonomous), Thanjavur, Tamilnadu, INDIA

Available online at: www.isca.in, www.isca.me
Received 18th July 2015, revised 18th August 2015, accepted 9th September 2015

Abstract

Monetary gauging or uniquely securities exchange expectation is one of the most blazing field of examination of late because of its profitmaking applications inferable from high stakes and the sorts of alluring advantages that it brings to the table. This paper introduces rough sets creating forecast guidelines plan for stock value development. The plan had the capacity separate information as principles from every day stock developments. These tenets formerly could be utilized to guide financial specialists whether to purchase, offer or hold a stock. Toward expand the effectiveness of the forecast procedure, rough sets with Boolean thinking discretization calculation is utilized to discretize the information. Rough set decrease method is connected to find every one of the reducts of the information. At long last, rough sets reliance guidelines are created specifically from every produced reduct. Harsh perplexity grid is utilized to assess the execution of the anticipated reducts and classes. The consequences of rough sets utilizing reducts structure by disarray network in choice table show general higher exactness rates of Decision making coming to more than 97% and create more minimized principle.

Keywords: stock market, stock market data prediction, Rough set theory, RS Algorithm.

Introduction

The stock investment has become an important daily part of life, however, the profits and risks of stock investment is directly related to each other. That means the higher investment income, the risks be greater[1].Therefore, there is an urgent need for an effective analysis method that can maximize returns and reduce risk. Common stock price forecasting methods: time series analysis, regression analysis, trend curve model method, Markov prediction method, discriminate analysis prediction method, etc. As the stock market have a lot of random factors, significantly affected the stock price, leading to price volatility, high noise, showing complex non-linear, uncertainty. Using traditional time series forecasting techniques is difficult to reveal its inherent laws, because the traditional methods are based on linear time series, these methods can’t fully take into account the price of non-linear characteristics, it can’t well analyze and fit the nonlinear relationship of predict stock, the prediction accuracy is low. With the development of nonlinear techniques, the rough set theory used in a wide range of stock market forecasts.

The rough set idea proposed by Pawlak[2] expect that there is some data which can be connected with each element of the universe. The same data can portray the items, and these are unintelligible in the perspective of accessible data about them. The supposed disjointed connection from this perspective is the numerical premise of the harsh set hypothesis. Rough set hypothesis[3] is likewise a numerical system that arrangements with dubiousness and vulnerability, and can be arranged inside of the fields of manmade brainpower (AI)[4], information revelation in databases and information mining (DM). Shen and Loh (2004)[5] did a detailed case study using the RS model to build a trading system in S and P 500 stock index. Their findings show that a RS model was an effective tool for forecasting S and P 500 stock index values. Jaaman et al.[6] investigated and forecast Malaysian stock market activities i.e. when to buy and sell a share by applying the RS methodology. Their results show that the RS model is an appropriate and operational method for stock market analysis. Recently Nair et al.[7] proposed a decision tree RS hybrid model for predicting the next day’s trend in Bombay stock exchange.

In this paper, we propose another standard based system, alleged rough set methodology, to focus business sector timing for securities exchange. Rough set methodology[8] is exceptionally significant to concentrate exchanging guidelines. To begin with, it doesn't make any suspicion about the conveyance of the information. Second, it can produce gainful business timing in light of the fact that it handles clamor well, as well as dispenses with insignificant variables[9]. Also, the harsh set methodology proper for distinguishing securities exchange timing in light of the fact that this methodology does not produce the sign for exchange when the example of business sector is dubious. Based on numerous studies conducted and their successful results, there is good reason and high probability that stock market prediction using the rough sets approach is applicable and promising. In the following, the utilization of the rough set theory to forecast the Nifty fifty Stock Exchange is focused. The rest of the paper is sorted out as takes after. In segment 2, we portray the a few meanings of rough set hypothesis.

In section 3, we describe the preprocessing of data and the discretization algorithms used. In section 4, the performance of the proposed approach is reported by experiment results. We will evaluate the discretization methods, the appropriate range of data use for induction, and the efficiency of the decision attributes to generated buy-hold-sell signals. Section 5 will conclude this study.

Preliminaries

We give below some definitions which will be used in this Section.

Information System

Data framework is of the form, a $tuple$ $(\mho, \mathcal{M})$, where ℧ contains of items and $\mathcal{M}$contains of elements. Each $\varpi \epsilon\, \mathcal{M}$ compares to the capacity $\varpi{:}\mho \rightarrow\ \gamma_{\varpi}$where $\gamma_{\varpi}$is a worth situated. Now, we usesfrequently recognize restrictive components A and choice elements B, where A∩ B =ϕ. In such cases, we characterize choice framework is (℧, A, B)

Indiscernibility Relation

Each subset of elements N⊆M impels incoherence connection

$$[Ind\ R]_N = [(m;\ n) \in \mho \times \mho : \forall\varpi \in \mathrm{N}; \varpi(m) = \varpi(n)\} \quad (1)$$

For each m∈℧, there is an equality class $X_B$in the segment of ℧ characterized by$[Ind\ R]_N$. Because of the roughness, which exists in certifiable information, there is now and again clashing arrangement of items contained in a choice table. The clashing arrangement happens at whatever point two articles have coordinating portrayals, yet are regarded to fit in with diverse choice classes. In such cases, the choice table is said to contain inconsistencies.

Lower and Upper Approximation

In Rough Set Theory, close estimations of sets are acquainted with manage irregularity. A harsh set approximates conventional sets utilizing a couple of sets named thelower and upper close estimation of the set. Given a setB ⊆ A, the lower and upper close estimations of a setY⊆℧ ; are characterized by comparisons (1) and (2) respectively.

$$\underline{BY} = \bigcup_{x:[x]_B \subseteq X} [m]_B \quad (2)$$

$$\overline{BY} = \bigcup_{x:[x]_{B\cap X \neq \phi}} [m]_B \quad (3)$$

Lower Approximation and positive region

The positive region $\mathrm{POS_C}$(D) is defined by

$$\mathrm{POS_C(D)} = \bigcup_{\mathrm{X:}\frac{\mathrm{X\epsilon U}}{\mathrm{Ind_D}}} \underline{\mathrm{C}}\mathrm{X}\ , \quad (4)$$

$\mathrm{POS_C(D)}$) is called the positive region of the partition $\mho/\mathrm{Ind\, R_D}$ with respect to $\mathrm{C} \subseteq\ \mathrm{A}$, i.e., the set of all elements in ℧ that can be distinctively classified by elementary sets in the partition $\mho/\mathrm{Ind\, R_D}$by means of C.

Upper Approximation and Negative Region

The negative region $\mathrm{NEG_C}$(D) is well-defined by

$$NEG_C(D) = U - \bigcup_{X:X\in\frac{U}{Ind\ R_D}} \overline{C}X \quad (5)$$

i.e., the set of all elements that can be definitely ruled out as members of X.

Boundary region

The limit locale is the distinction in the middle of upper and lower estimate of a set X that comprises of equality classes taking one or more components in the same manner as X. It is given as takes after:

$$\mathrm{BND_B(X)} = \ \underline{\mathrm{B}}\mathrm{X} - \overline{\mathrm{B}}\mathrm{X} \quad (6)$$

Reduct

Givenagrouping errand identified withthemapping C→ D, a reduct is a subset R ⊆ C such that γ(C, D) =γ(R, D) and none of fitting subsets of R fulfills comparable to uniform.

Reduct Set

Given a grouping undertaking mapping an arrangement of variables C to an arrangement of marking to an arrangement of naming D; a reduct set is characterized regarding the force set P(C) as the set R ⊆ P(C) such that

$$\mathrm{Red}\ \ = \{\mathrm{A} \in \mathrm{P(C)}: \gamma(\mathrm{A, D})\}. \quad (7)$$

That is, the reduct set is the set of all possible reducts of the equivalence relation denoted by C and D.

Minimal Reduct

A minimal reduct$\mathrm{R_{minimal}}$ is the reduct such that

$$\|\mathrm{R}\| \leq \|\mathrm{A}\|, \forall \mathrm{A} \in \mathrm{R}. \quad (8)$$

That is, the minimal reduct is the reduct of least cardinality for the equivalence relation denoted by C and D.

Core

Trait c ∈C is a center element as for D, if and on the off chance that it fits in with every one of the reducts. We signify the arrangement of all center elements by Core(C). On the off chance that we signify by R(C) the arrangement of all reducts, we can put:

$$Core(C) = \bigcap_{R\in R(C)} R \quad (9)$$

Significance

For any feature$\mathrm{a} \in \mathrm{C}$, we define its significance ζ with respect to D as follows:

$$\zeta(a, C, D) = \frac{|POS_{C\backslash\{a\}}(D)|}{|POS_C(D)|} \quad (10)$$

Algorithms

Rough Set Algorithm 1 (Information table (ST) with discretized real valued attribute**):**

```
Input: Information framework table (S) with genuine esteemed traits $A_{ij}$and n is the quantity of interims for every quality in ITC Dataset.

Output: Information table (ST) with discretized genuine esteemed characteristic in ITC Dataset.
for$A_{ij} \in S$do
    Describe a set of Boolean variables as follows:$B = \{\sum_{i=1}^{N} C_{ai}, \sum_{i=1}^{N} C_{bi}, \sum_{i=1}^{N} C_{ci}, \dots\dots, \sum_{i=1}^{N} C_{Ni}, \}$    (11)
    Where$\sum_{i=1}^{N} C_{ai}$,relate to a set of intervals well-defined on the variables of elements a.
end for
Create a new information table $S_{new}$ by using the set of intervals $C_{ai}$
Find the minimal subset of $C_{ai}$ that discerns all the elements in the decision class D using the resulting formula:
    $Y^u = \wedge\{\Phi(i, j) : d(x_i) \neq d(x_j)\}$    (12)
Where, $\Phi(i, j)$is the number of minimal cuts that must be used to discern two differentInstances $x_i$ and $x_j$ in the information table.
```

Rough Set Algorithm 2 (Reduct Sets**):**

```
Input: information table (ST) with discretized real valued attribute from ITC Dataset.
Output: reduct sets detection in ITC Dataset where as$R_{final} = \{r_1 \cup r_2 \cup \dots\dots \cup r_n\}$
1:  for each condition elements$c \in C$do
2:      Calculate correlation factor between c and the decisions elements D
3:      ifthe $correlation\ factor > 0$ then
4:          Set $c$ as related elements.
5:      end if
6:  end for
7:  Divide the set of related element into different variable sets.
8:  for each variable sets do
9:      Calculate the dependency degree and calculate the classification value
10:     Let the set with high classification accuracy and high dependency or support as an first reduct set.
11: end for
12: for each element in the reduct set do
13:     Calculate the degree of dependencies between the decisions attribute and that element.
14:     Combine the elements produced in earlier step with the rest of conditional elements
15:     Calculate the discrimination factors for each combination to find the highest discrimination factors
16:     Add the highest discrimination factors combination to final reduct set.
17: end for
18: repeat Statements 12
19: until all elements in initial reduct set is processed.
```

Rough Set Algorithm 3 (Set of rules**):**

```
Input: reduct sets in ITC Dataset$R_{final} = \{r_1 \cup r_2 \cup \ldots . . \cup r_n\}$
Output: Set of rules framed based on Change in Support and Accuracy in ITC Dataset.
$\boldsymbol{for}$ each reduct $r$ do
    $\boldsymbol{for}$ each communication item $x$ $\boldsymbol{do}$
        Contract decision rule $(c_1 = v_1 \wedge c_2 = v_2 \wedge \ldots . . \wedge c_n = v_n) \rightarrow d = u$
        Scan the reduct$r$ over an item $x$
        Construct $(c_i, 1 \leq i \leq n)$
        $\boldsymbol{for}$every $c \in C$ $\boldsymbol{do}$
            Allocate the value v to the correspondence element$a$
        $\boldsymbol{end\ for}$
        Create a decision element$d$
        Allocate the value u to the correspondence decision element$d$
    $\boldsymbol{end\ for}$
$\boldsymbol{end\ for}$
```

Results

The factual investigation in rough Set is utilized to speak to vital conveyance of traits, in view of this representation serves to achieve the negligible number of reducts that has a blend of characteristics which has the similar segregation variable. The last created reduct sets which are utilized to produce the rundown of guidelines for the characterization are:

Information speaking to the end value, opening value, most astounding value came to amid the day, and the least value came to amid the day.

The Rough set guidelines are utilized to remove the test dataset of ITC Company from the preparation dataset Nifty 50 Companies.

The preprocessed information was part into examination of preparing arrangements of 3145 articles for information dataset and approval testing specimens of 251 items sets for ITC Company. The investigation set was used by harsh sets, which obtained reducts and choice principles.

Every choice standard had estimations of bolster, exactness, and scope (table-1). These guidelines be situated the essential yield of the rough set investigation method. The choice standards were formerly approved by terminating every specific principle in conjunction with the acceptance information set. The new bolster, exactness, and scope measures were watched for the acceptance information set. The objective is to discover principles that are exact representations of the information. Along these lines, decides that have comparable measures in both the investigation and approval sets ought to existmeasuredby way of steady then exact.

The situation is intriguing to tell the consequence of consolidating two choice classes. Two conceivable blends can be achieved: the mix of a diminishing and unbiased, and the mix of an increment and impartial. The mix of diminishing and build conceivably offers no quality. Notwithstanding, the mix of nonpartisan and either build or decline can let us know whether this standard can offer a huge addition or misfortune with the incorporation of a peripheral change. Principle 1 is a case of a standard with estimations got through both the examination and acceptance sets.

Every standard is connected to the approval information set relating to that of the investigation set. Data is gathered with respect to the backing and precision of that control in the new information. The four decides that were gained through the procedure cover an aggregate 3145 of 251 objects of High Price, and 3145 and 251 objects of Low cost in the examination set and approval set individually.

Table-1
Indicates Sample Testing Dataset of ITC Company

Date	Open Price Value	High Price Value	Low Price Value	Close Price Value	Total Volume Value	No of Trades	Turnover in (Rs.in Lakh) Value
9/11/2014	353	349.5	351.1	351.2	58,74,927	79,236	2,06,37,78,770.00
9/10/2014	357	350.8	357	351.05	67,39,816	85,441	2,37,83,50,377.00
9/9/2014	359.6	351	351	358.2	38,62,825	45,442	1,37,58,30,388.00
9/8/2014	355	350.3	353	353.95	42,13,391	39,431	1,48,76,10,575.00
9/5/2014	352	348.4	350.6	350.8	48,73,614	76,502	1,70,80,16,871.00
9/4/2014	352.9	349.1	349.6	350.4	58,59,623	58,901	2,05,41,61,220.00
9/3/2014	355.7	348.3	355.3	348.9	65,50,532	65,857	2,30,04,60,415.00
9/2/2014	356.2	349	350.5	354.7	48,90,327	60,277	1,72,50,65,206.00
9/1/2014	358	349.7	358	350.55	47,12,322	80,702	1,66,42,83,548.00
8/28/2014	357	351.9	352	355.3	84,09,155	68,855	2,98,12,06,816.00

Table-2
Statistical Analysis of Input Training Dataset of ITC Company

Parameters	High Prices		Low Prices	
Dataset	Analysis Set(Training)	Validation Set(Test)	Analysis Set(Training)	Validation Set(Test)
Total Objects	144315	3145	144315	3145
Objects Covered	3145	251	3145	251
Min Support	117	157.6	104.65	154.3
Max Support	2580	368.25	2561	364.8
Average Support	586.827217	245.0029101	575.993815	241.4039683
Min Accuracy	0.003794232	0.003598894	0.00984204	0.010014306
Max Accuracy	0.009109542	0.006348149	0.023526365	0.017526593
Average Accuracy	0.005245882	0.005254391	0.013731128	0.014588076

Table-3
Sample Dataset ITC Company for Rough Set Rule Prediction and Ranking
Number of clusters selected by cross validation:7

Attribute		**Clusters**						
		0	**1**	**2**	**3**	**4**	**5**	**6**
		(0.2)	**(0.12)**	**(0.05)**	**(0.14)**	**(0.12)**	**(0.17)**	**(0.21)**
Open price	Mean	347.944	355.9522	297.1681	327.8815	307.6196	337.1521	319.5516
	Std.dev	3.4563	2.3836	3.7589	3.1564	4.7436	2.952	2.767
High Price	Mean	340.1732	348.6746	288.1945	320.6939	300.426	330.3475	312.9774
	Std.dev	2.8859	3.1553	3.1557	3.0043	4.1066	2.7269	3.0166
Low Price	Mean	344.7228	352.4892	293.3137	324.6902	304.842	333.8499	316.3103
	Std.dev	4.0819	3.634	3.3836	3.7507	4.8928	3.4622	3.1404
Close Price	Mean	343.7641	353.192	292.5575	324.3377	304.1623	334.1813	316.5406
	Std.dev	3.0367	3.3006	4.0791	3.0557	3.8435	2.8129	3.0169
Class	ITC	79.4816	47.2819	21.466	55.7648	45.7781	66.2976	59.6143
	Total	79.4816	47.2819	21.466	55.7648	45.7781	66.2976	59.6143

Table-4
Number of iterations performed: 16

Clustered	Instances
0	78 (20%)
1	46 (12%)
2	21 (5%)
3	53 (14%)
4	44 (11%)
5	66 (17%)
6	60 (21%)

So as to conclude that which rules are be strongest, *one must* relate the other rules and in what way they react with data. On the way to investigate the rules that were have obtained, a ranking method could be used. Specific rules are ranked according to their accuracy and stability. Enchanting into account all different types of rankings that were used, an overall rank can be determined. The results of this method can be seen in table-3 where the header R_1, R_2 , R_3, R_4, R_5, R_6, R_7 and R_8 are rankings according to total support, total accuracy, and change in support and change in accuracy respectively. The last rank is determined by the previous 8 rankings for each rule.

Table-5
Statistical Results of Attribute Values

Rule	Change in support (%)	Change in accuracy (%)
1	78	0.2
2	46	0.12
3	21	0.05
4	53	0.14
5	44	0.12
6	66	0.17
7	60	0.15
8	18	0.04

Table-6
Statistical Results of Attribute Values

Rule	Change in accuracy (%)	Rank
8	0.04	1
3	0.05	2
2	0.12	3
5	0.12	4
4	0.14	5
7	0.15	6
6	0.17	7
1	0.2	8

Table-7
Decision Table: Ranking of Rules (Lower is better)

Rule 8 → Change in Accuracy varies with 0.04 less than or equal to 0.05 in ITC Clustered Data is ranked as No 1.
Rule 3→ Change in Accuracy varies with 0.05less than or equal to 0.12 in ITC Clustered Data is ranked as No 2.
Rule 2→ Change in Accuracy varies with 0.12less than or equal to 0.14 in ITC Clustered Data is ranked as No 3.
Rule 5→ Change in Accuracy varies with 0.12 less than or equal to 0.14 in ITC Clustered Data is ranked as No 4
Rule 4→ Change in Accuracy varies with 0.14 less than or equal to 0.15 in ITC Clustered Data is ranked as No 5.
Rule 7→ Change in Accuracy varies with 0.15less than or equal to 0.17 in ITC Clustered Data is ranked as No 6.
Rule 6→ Change in Accuracy varies with 0.17less than or equal to 0.2 in ITC Clustered Data is ranked as No 7.
Rule 1→ Change in Accuracy varies with 0.2and above in ITC Clustered Data is ranked as No 8.

The accuracy of ranking value shown in table-6, Rule 8 is recommended by way of the best with low rankings for measures in support, accuracy, change in support, and change in accuracy. Rule 1 is measured the poorest of the eight rules based on the high rankings of total support and total accuracy. Rules 6 and 7 both have a better ranking after Rule1.

This is due to the point that rule 3 takes high rankings for total support and change in accuracy but low rankings for accuracy and change in support. Rule 2 and 5 takes average rankings for each measure.

A consumer can provide different weights to different rankings. For example, if a consumer desires to rank rules according to stability, he or she may combine a higher level of consideration 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 provided to each measure.

The figure-1 indicates the Rough Set Prediction of Low Price Value and High Price Value in Stock Market Data from Jan

2008 to Sep 2014. The Sample Data is used to detect reducts using Rough Set and predict the High Price and Low Price value for ITC Company.

The figure-2 indicates the Rough Set Prediction of High Price Value in Stock Market Data from Jan 2008 to Sep 2014. The Sample Data is used to detect reducts using Rough Set and predict the High Price value for ITC Company.

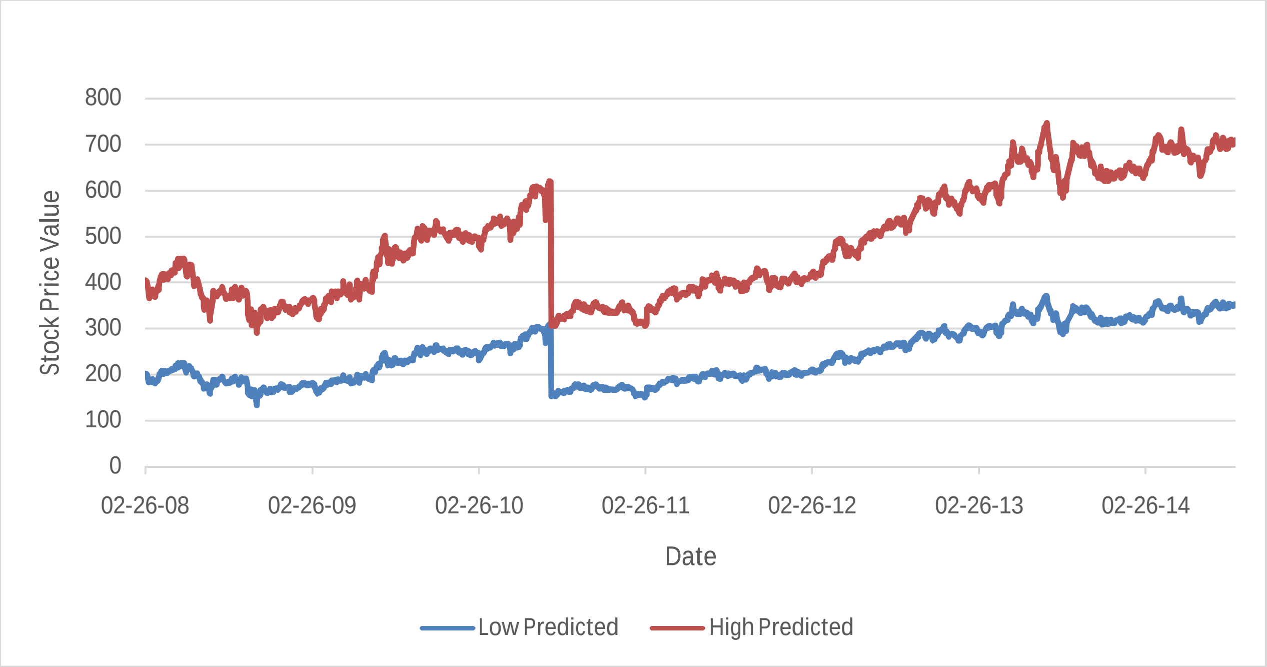

Figure-1
Rough Set Prediction of Low Price Value and High Price Value in ITC Company

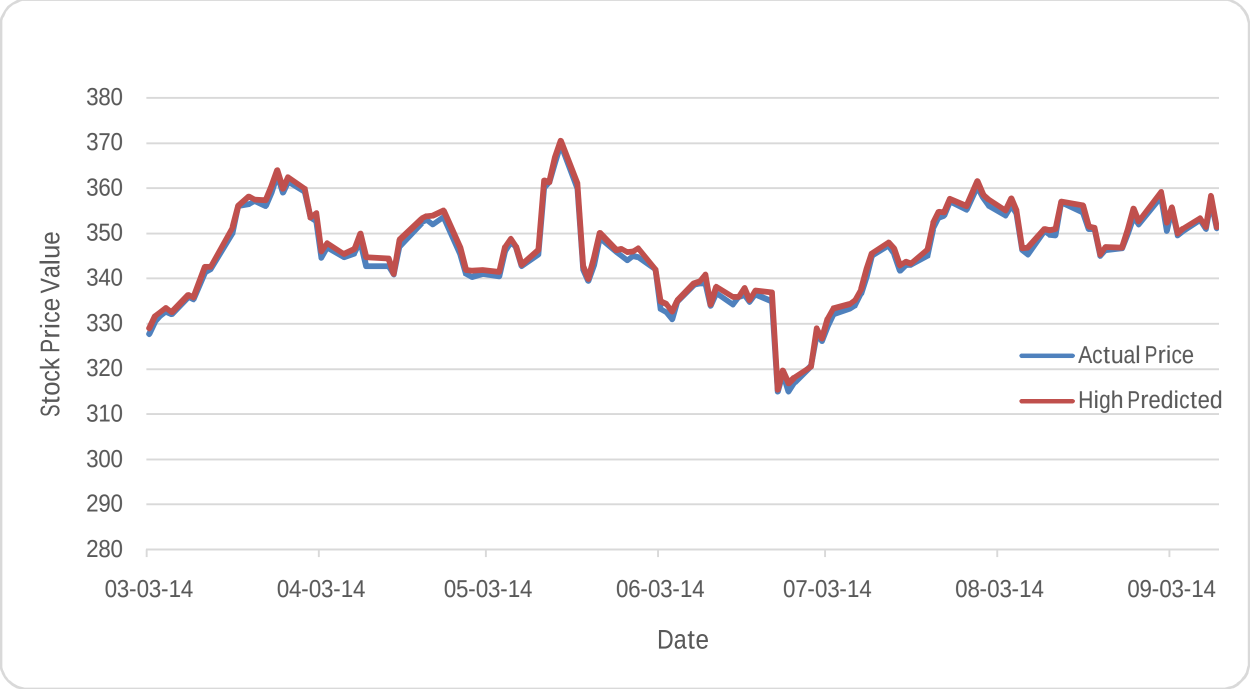

Figure-2
Rough Set Prediction of High Price Value in ITC Company

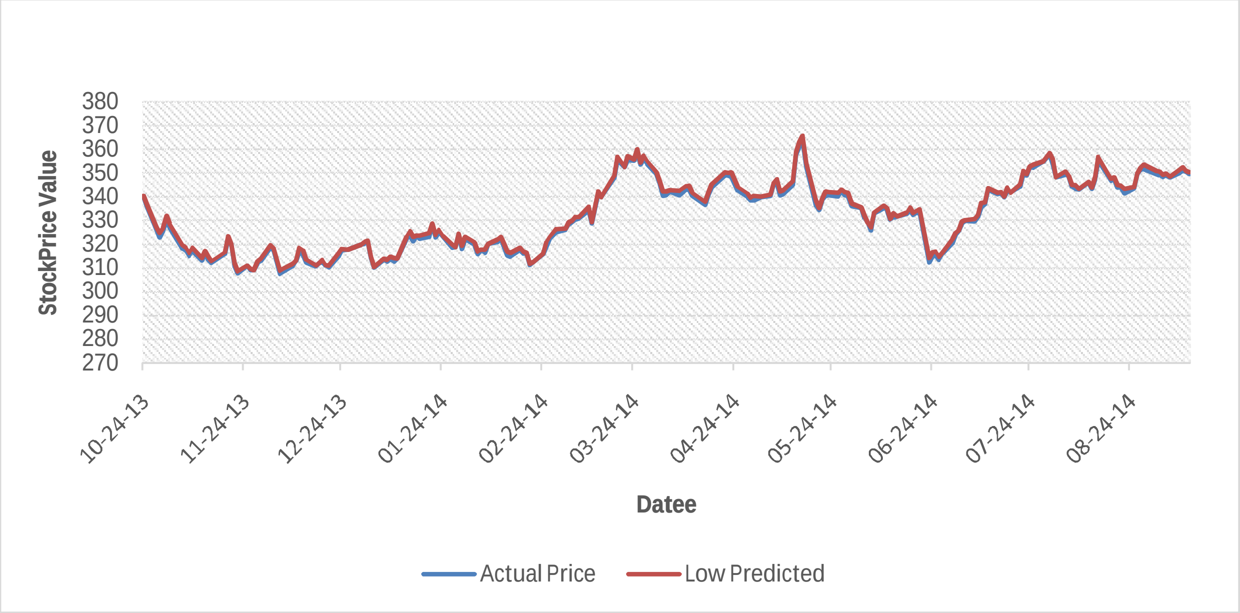

Figure-3
Rough Set Prediction of LowPrice Value in ITC Company

The figure 3 indicates the Rough Set Prediction of Low Price Value in Stock Market Data from Jan 2008 to Sep 2014. The Sample Data is used to detect reducts using Rough Set and predict the Low value for ITC Company.

Conclusion

In the above review of rough sets model for economic and financial prediction, it has been demonstrated that rough sets model is a promising alternative method to conventional methods. However, thus far the use of Rough Sets Theory has been restricted to the classification problem because classification objects can be directly put into the decision table. This usage can be generalized from the applications described in previous sections. The usage of rough sets theory to the choice and ranking problems, which are often encountered in economic and financial decision making problems. In their approaches, the conventional decision table is replaced by a pairwise comparison table, i.e., a decision table whose objects and entries is pairs of actions and binary relation instead of single action and attributes values, respectively. In this way, the built preferential model is much closer to the natural reasoning of the decision making problem. This approach will broaden the application of Rough Sets Theory in the economic and financial problems.

References

1. Yang xinbin and Huang xiaojua, Study about Application of Stock Price Forecasting Based on Support Vector Machine [J],. *Computer Simulation,* **27(9),** 302-305, **(2010)**

2. Pawlak Z., Rough Sets, *International Journal of Computer and Information Science,* 341-356, **(1982)**

3. Pawlak Z., Rough Sets: Theoretical Aspects of Reasoning about Data, System Theory, Knowledge Engineering and Problem Solving, vol. 9, Kluwer Academic Publishers, Dordrecht, The Netherlands, **(1991)**

4. E. Gately, Neural Networks for Financial Forecasting; Wiley: New York, NY, USA, **(1996)**

5. L. Shen and H.T. Loh, Applying Rough Set to Market Decisions, *Decisions Support Systems,* **37,** 583-597**, (2004)**

6. S.H. Jaaman, S.M. Shamsuddin, B Yusob and M. Ismail, A Predictive Model Construction Applying Rough Set Methodology for Malaysian Stock Market Returns, *International Research Journal of Finance and Economics,* 211-218**, (2009)**

7. B.B Nair, V.P Mohandas and N.R. Sakthivel, A Decision Tree Rough Set Hybrid System for Stock Market Trend Prediction, *International Journal of Computer Applications,* **.6,** 1-6, **(2010)**

8. Pawlak Z., Rough Relations, Reports, vol. 435, Institute of Computer Science, Polish Academy of Sciences Warsaw, Poland, **(1981)**

9. Ruggiero M.A., Cybernetic Trading Strategies: Developing Profitable Trading Systems with Stateof-the-art Technologies, John Wiley and Sons, Inc, **(1997**)