Predictive Data Mining in KPP

Anupam Bhatia [#1], Dr. R.K. Chauhan [*2]
[#] *Assistant Professor, Kurukshetra University Post Graduate Regional Centre, Jind (India)*
[1] *anupam.bhatia@kuk.ac.in*
[*] *Professor, Department of Computer Science and Applications, Kurukshetra University, Kurukshetra (India)*
[2] rkchauhankuk@gmail.com

Abstract— In this paper, we have provided the Genetic Algorithm (GA) used for prediction process in Knowledge Penetration Process (KPP). The said GA is implemented and its efficiency is analyzed.

Keywords: Knowledge Penetration Process, Predictive Data Mining, Prediction, Genetic Algorithm

I. INTRODUCTION

It is the step of Mining phase of KPP that performs inference on the current data in order to make predictions. In our research, this step is referred as *Prediction*.

II. GENETIC ALGORITHM FOR PREDICTION

There are more reasons for preference using genetic programming and genetic algorithms in general in contrast to other techniques. One of them is their robustness and ability to work on large and "noisy" datasets, they perform global search of the solution space in comparison to most other algorithms that use greedy search, coping well with attribute interaction. Owing to all possible modifications and parallel approaches to genetic algorithms, the scalability of these algorithms can be achieved. Beside robustness, this characteristic is of great importance in data mining. Moreover, these algorithms have high degree of autonomy that enables discovery of knowledge previously unknown by the user. The problem of comprehensibility of discovered rules can be addressed by properly adjusting the fitness function.

1. Start
2. Initialize the Population
3. Initialize the program size
4. Define the fitness f_i of an individual program corresponds to the number of hits and is evaluated by the formula:

```
zhat = predict(z, se.fit=TRUE)
zupper =zhat$fit + 1.96 * zhat$se.fit
zlower = zhat$fit - 1.96 * zhat$se.fit
yupper = exp(zupper)/(1 + exp(zupper))
ylower = exp(zlower)/(1 + exp(zlower))
```

5. Run a tournament to compare four programs randomly out of the population of programs
6. Compare them and pick two winners and two losers based on fitness

7. a) Copy the two winners and replace the losers
   b) With Crossover frequency, crossover the copies of the winners
   c) With Mutation frequency, mutate the one of the programs resulting from performing step 7(a)
   d) With Mutation frequency, mutate the other of the programs resulting from performing step 7(a)

Repeat through step 5 till termination criteria are matched.

III. IMPLEMTATION

All paragraphs must be indented. All paragraphs must be justified, i.e. both left-justified and right-justified.

A. Data Set

The National Stock Exchange is a stock exchange located at Mumbai, Maharashtra, India. It is the 16th largest stock exchange in the world by market capitalization and largest in India by daily turnover and number of trades, for both equities and derivative trading. The subset of data set of NSE for the financial year 2010-11 is used for implementation.

B. Data Mining Tools Used

For the preprocessing steps Notitia Data Preparation Software is used.
For the processing of Data Mining steps Discipulus Software is used.

C. Implementation

Original data set is available in Microsoft Excel Workbook. Each worksheet of the workbook contains 50 days data of NIFTY. Each worksheet is imported as a distinct data set for the Data Mining processing purpose.

First the data is imported from selected worksheet. In the Raw Data tab Cleaning Process is performed by filling up the missing values or excluding the missing value columns and handling the outliers. Columns not required for specific mining process are also excluded.

As soon as the data is cleaned the included columns are checked for transformation purpose. In our research, we have used two different columns as output. One is named "out" for prediction purpose and other is "out1" for classification purpose. In each processing step output column is selected and transformed accordingly

In prediction, it is predicted on the basis of 50 days data whether on the 51st day the value of share will increase or decrease. If it will increase, it is labeled "1" otherwise it is labeled "0".

Data Split

It is essential to split the data set for training and test purpose. We have split our data set in three parts

1. Training Data Set:- Training Data Set is subset of data used for training purpose.
2. Validation Data Set:- Validation Data Set is another subset of data used for validation purpose of training. It is simply a test set which is mutually exclusive to training data set.
3. Applied Data Set:- Applied Data Set is mutually exclusive to training and validation data set. It is either subset of data set used for training and validation or new data set with same attributes

For the prediction, we have implemented Genetic Algorithm as stated in Section 2.

The crieteria of termination is;

Generation Without Improvement = 700
Project Termination after number of runs = 100
To check the quality of Prediction output, ROC charts are generated and interpreted.

In ROC charts, Area Under Curve is taken as quality measure.

Data Set 1

In data set 1 data is split in three equal size disjoint data sets.

First part is used as Training Data Set. It is one third part of Data Set 1. In rest of all the computations, same data set and its results are used as training data set as intermediate results, auxiliary data, auxiliary statistics and auxiliary tuple.
Second part is used as Validation Data Set. In rest of all the computations, same data set and its results are used as validation data set as intermediate results, auxiliary data, auxiliary statistics and auxiliary tuple.
Third part is used as Applied Data Set. It is one third part of data set 1
All splitted data sets i.e. training, validation and applied are disjoint to each other.

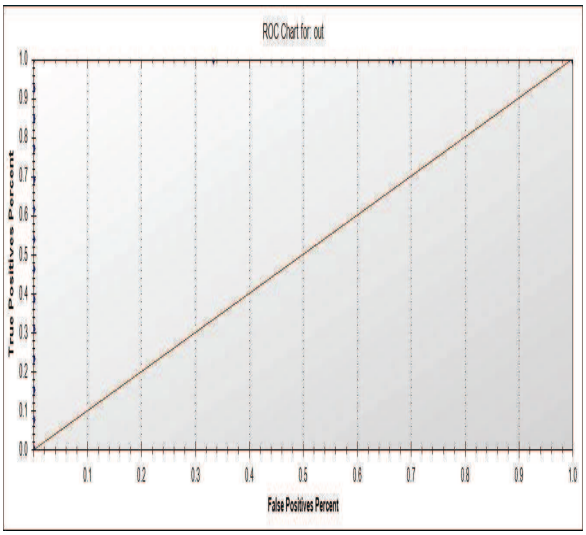


Fig 1: ROC Chart for Training Data Set

The ROC Chart as shown in Fig. 1 is generated with Training Data Set.

Area Under Curve is found 1.

It is a well known fact that in ROC Chart, the closer the area is to 0.5, the less accurate the corresponding model is. A model with perfect accuracy have an area under curve is 1.0.

From the data of highly volatile market of National Stock Exchange, India; candidate data set with perfect training accuracy is significant, motivational compatible and outstanding.

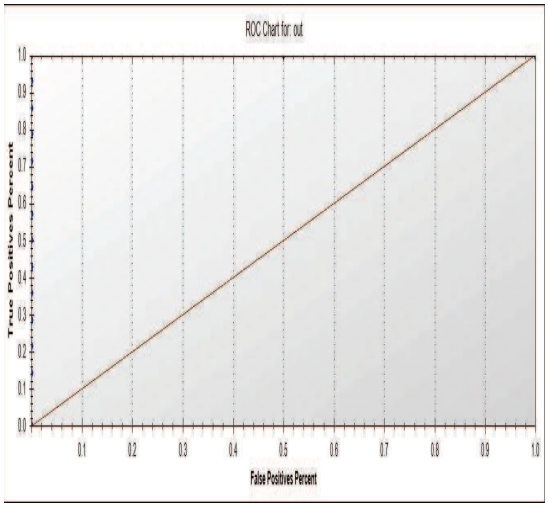


Fig 2: ROC Chart for Validation Data Set

The ROC Chart as shown in Fig.2 is generated with Validation Data Set.

Area Under Curve is found to be 1.

It is a well known fact that in ROC Chart, the closer the area is to 0.5, the less accurate the corresponding model is. A model with perfect accuracy have an area under curve is 1.0.

From the data of highly volatile market of National Stock Exchange, India; validation of candidate data set with perfect accuracy is significant, motivational compatible and outstanding.

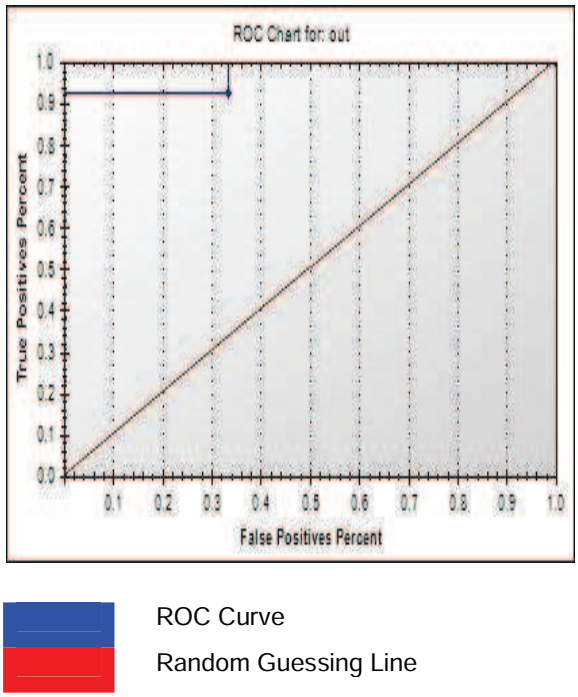


Fig 3: ROC Chart for Applied Data Set1

The ROC Chart as shown in Fig.3 is generated with Applied Data Set1.

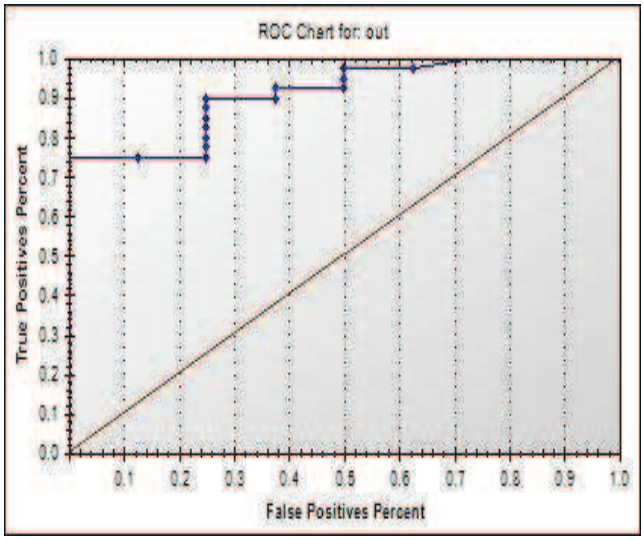
Area Under Curve is found 0.9743 which is very close to 1.

It is a well known fact that in ROC Chart, the closer the area is to 0.5, the less accurate the corresponding model is. A model with perfect accuracy have an area under curve is 1.0.

From the data of highly volatile market of National Stock Exchange, India; candidate data set with almost perfect applied accuracy is significant, motivational compatible, outstanding and satisfy user interestingness.

Data Set 2

In data set 2, training and validation data from data set 1 is used. Whole data set 3 is used as applied data set.



ROC Curve
Random Guessing Line

Fig 4: ROC Chart for Applied Data Set 2

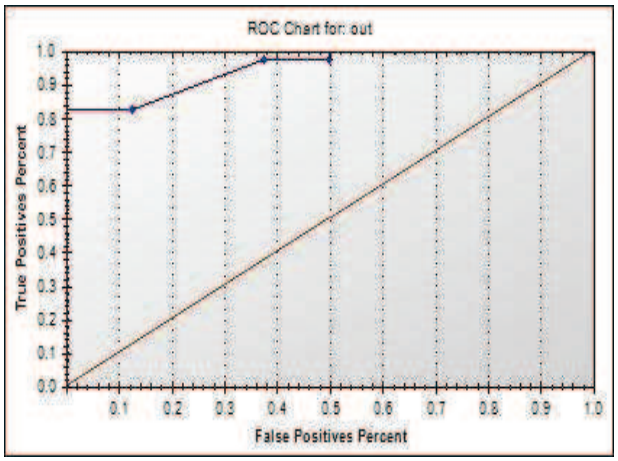
The ROC Chart as shown in Fig. 4 is generated with Applied Data Set 2.
Area Under Curve is found 0.9109 which is close to 1.

It is a well known fact that in ROC Chart, the closer the area is to 0.5, the less accurate the corresponding model is. A model with perfect accuracy have an area under curve is 1.0.

From the data of highly volatile market of National Stock Exchange, India; candidate data set with almost perfect applied accuracy is significant, motivational compatible, outstanding and satisfy user interestingness.

Data Set 3

In data set 3, training and validation data from data set 1 is used. Whole data set 3 is used as applied data set.



ROC Curve
Random Guessing Line

Fig 5: ROC Chart for Applied Data Set3

The ROC Chart as shown in Fig. 5 is generated with Applied Data Set3.
Area Under Curve is found 0.95 which is very close to 1.
It is a well known fact that in ROC Chart, the closer the area is to 0.5, the less accurate the corresponding model is. A model with perfect accuracy have an area under curve is 1.0.

From the data of highly volatile market of National Stock Exchange, India; candidate data set with almost perfect applied accuracy is significant, motivational compatible, outstanding and satisfy user interestingness]

IV. CONCLUSION

In our research, we have worked on the improvement of existing KDD process. For the purpose a specific model named as Knowledge Penetration Process (KPP) is designed and explained. KPP is a five step model consists of phases *Problem Analysis, Preprocessing, Pre Mining, Mining and Evaluation.* To improve the quality of knowledge, Genetic Algorithms are used in the core mining process of KPP. To

improve the runtime, intermediate results, auxiliary data, auxiliary tuples and auxiliary statistics are used.

To check the efficiency of the model, various processes are implemented by using the data set of National Stock Exchange of India. KPP model is implemented three times on different data sets. In the Evaluation phase of KPP, the quality of results is measured.

The quality of *Prediction* is measured with *ROC Curve*.

Training and Validation are repeated until 100% accuracy is achieved in Prediction.

From the data of highly volatile market of National Stock Exchange, India; candidate data set with almost perfect applied accuracy is significant, motivational compatible, outstanding and satisfy user interestingness.

REFERENCES

[1] *Lin, Li. Cao, Longbing. et. al, The Applications of Genetic Algorithms in Stock Market Data Mining Optimisation, Capital Market CRC, Sydney NSW, Australia , 2000*

[2] *Chauhan R.K., Bhatia Anupam, KPP : A Step Ahead to KDD, RIMT Journal of Strategic Management and Information Technology , pp. 174-177, 2008*

[3] *Weber, Ben G. Mateas, Michael (2009). "A Data Mining Approach to Strategy Prediction" 978-1-4244-4815 2009 IEEE.*

[4] *Bhatia Anupam, Chauhan R.K., Knowledge Penetration Process, A Splitted KDD, Global Journal of Computer Science and Technology, USA, 2011*

[5] *Olaniyi, S Abdulsalam. Kayode, S.,(2011). "Stock Trend Prediction Using Regression Analysis – A Data Mining Approach". ARPN Journal of Systems and Software, Volume 1 No. 4.*

[6] *http://www.nseindia.com/content/indices/ind_niftylist.csv*

[7] *www.rmltech.com*