An Association Rule Mining based Stock Market Recommender system

Preeti Paranjape-Voditel
Department of Computer Applications,
Shri Ramdeobaba Kamla Nehru Engineering College,
Nagpur, India
e-mail: preetivoditel@gmail.com

Dr.Umesh Deshpande
Department of Computer Science and Engineering,
Visvesvarayya National Institute of Technology,
Nagpur, India
email: uadeshpande@gmail.com

Abstract— **We propose an Association Rule Mining (ARM) based Recommender system for the stock markets. Normally technical and fundamental analyses are a basis of prediction of stock price. Several systems exist for monitoring and prediction of stock prices. But these deal with individual stocks. They do not give the inter-relationship between stocks or their relationship with the stock market INDEX. Our method uses ARM, fuzzy ARM, weighted fuzzy ARM, ARM with time lags, fuzzy ARM with time lags and weighted fuzzy ARM with time lags to predict relationships between stocks, which is used as the basis for portfolio management and in recommendations for mutual funds.**

Keywords - recommender system, fuzzy ARM, time lags, fuzzy time lagged ARM, weighted fuzzy ARM, weighted fuzzy time-lagged ARM

I. INTRODUCTION

Recommender systems are software applications that aim to support users in their decision making, where large information spaces are concerned. They effectively prune large information spaces and direct the users towards the items that best meet their needs and preferences. Such systems have an obvious appeal in an environment where the amount of on-line information vastly outstrips any individual's capability to survey it [4].

A Recommender system works from a specific type of information filtering system that attempts to recommend items that can of interest to the user. These information items can be films, television shows, books, images, web pages, music scores, stocks etc. Typically, a recommender system compares a user profile to some reference characteristics, and seeks to predict the 'rating' that a user would give to an item he had not yet considered. These characteristics may be from the information item (the content-based approach) or the user's social environment (the collaborative filtering approach).

Typical applications of recommender systems are product finders, stock market predictors, shopping guides, etc.

Ratings are not always provided and that is when item based collaborative filtering of data such as purchase data, stock market data, weblog data can be used.

Stock prices depend on various factors, the important ones being the market sentiment, performance of the industry, earning results and projected earnings, takeover or merger, introduction of a new product or introduction of an existing product into new markets, share buy-back, announcements of dividends/bonuses, addition or removal from the index and such other factors leading to a positive or negative impact on the share price and the associated volumes.

Apart from the basic technical and fundamental analysis techniques used in stock market analysis and prediction, soft computing methods based on Association Rule Mining, fuzzy logic, neural networks, genetic algorithms etc. are increasingly finding their place in understanding and predicting the financial markets.

Association rule discovery has been used with great success in domains such as market basket analysis but it finds an even wider domain of applications when used in combination with other classification and predictive approaches. We have used Association Rule Mining along with fuzzy classification methods to develop our Recommender system for the stock markets. It can be used for dynamic portfolio management and also in mutual funds.

The organization of the paper is as follows: Section II deals with the existing work in this domain, Section III describes the Recommender system, Section IV describes the fuzzy association rule generation, Section V deals with the weighted fuzzy association rule mining, Section VI describes weighted fuzzy time lagged association rule generation.

II. EXISTING WORK IN THIS DOMAIN

Statistical methods for individual stocks like the Newton method [10], logarithmic regret algorithms [11] etc have been used widely for portfolio management. But they deal with the price and volume of the stocks in calculating the gains or the losses for the portfolio.

A good amount of work has been done on predicting stock prices for individual stocks [6,7,8,9]. In [5], the FP-tree based frequent itemset mining to mine inter-transaction association rules is used. The number of possible rules are limited using a sliding window. The Fuzzy logic based stock Trading System described in [3], used by the "KOLEGU" mutual fund applies certain constraints based on the fundamentals of the stock and then generates rules/recommendations for buying or selling. [2] uses temporal data mining to build the Intelligent stock market assistant capable of suggesting to buy or sell stocks. It also takes into consideration the risk level, buying price and the need for cash.

978-0-7695-4329-1/11 $26.00 © 2011 IEEE
DOI 10.1109/EAIT.2011.90

In our Recommender System, the basic underlying theory is the generation of Association Rules. The Association Rules generated in our system are divided into the following categories: 1)Intra-sector 2)Inter-sector 3)Time-lagged 4) Fuzzy 5)Fuzzy time-lagged 6)Weighted fuzzy 7) weighted fuzzy time-lagged. The Recommender System handles inter-day as well as intra-day associations. The generation of these rules and their role in recommendations will be discussed in the following sections.

III. DESCRIPTION OF THE RECOMMENDER SYSTEM

The system mines relationships between items or scrips, in our case, based on a support- confidence framework. The system can be used for portfolio management where we assume an existing portfolio and it has to be managed with alternative replacement for scrips. It does not recommend the scrips in isolation but in relation to the other existing scrips. The objective is to show good returns. The user can input his own portfolio or if a portfolio has to be created, it is done by taking the amount to be invested from the user as input and one scrip from each sector is invested in.

The transaction files for this system were created by finding out the percentage rise/fall of a certain scrip from its previous trading day's close. This was done for the 30-scrip BSE SENSEX which comprises of the blue chip, widely traded companies of the Bombay Stock Exchange. Thus a transaction will contain all the scrips which have risen/fallen by more than some minimum amount. The above files were created for data for the past five years.

B. Generation of the Datasets

Our dataset consists of the Bombay Stock Exchange (BSE) Sensitive Index called the SENSEX scrips as the BSE-SENSEX is the barometer of the market behavior in the Indian markets. Also this can be easily extended to the National Stock Exchange (NSE) scrips.

The source of price-volume data is taken from the official End-Of-Day (EOD) Quotes for BSE India Stocks and Indices. It contains the BSE SENSEX information like the daily opening and closing price associated with the volume. From this data our dataset has been created. Each transaction in this dataset consists of all those scrips which have risen by a particular percentage over the previous close. This parameter can be defined by the user or by default this value is taken between 1 and 2% by our algorithm depending upon the type of dataset to be generated.

Stock markets, in India or in other countries, can be divided into sectors and the inter-sector and intra-sector relationships can be found. We have divided the SENSEX scrips into seven sectors according to their area of operation and also according to their weightage in the SENSEX as shown in Table1. The Inter-sector relationships between them help in understanding as to which sectors are independent of each other and which sectors have a precedent-antecedent relationship in terms of price rise/fall. In the Recommender system, if a sector scrip exists in the portfolio and the sector faces a slump, a negatively correlated scrip can be included. If a sector performs phenomenally well, more positively correlated scrips from that sector can be included.

SEC NO.	SCRIPS IN THE SECTOR	SECTOR	APPROXIMATE WEIGHTS AS PER SENSEX
1.	ACC, BHEL, DLF, Jaiprakash Ind,L&T, Tata Steel	Cement, Engineering, Construction, Steel	0.7+4.43+2.02+1.23+3.71+2.18=14.27
2.	ONGC, Reliance	Oil	8.61+13.65=29.31
3.	NTPC, Rel.Infra, Sterlite, Tata Power	Power, Metals	0.88+2.64+1.23=4.75
4.	Bharti Airtel, Infosys, Rel.Comm, TCS, Wipro	Telecom, Computers	4.51+6.06+1.32+6.12+4.06=22.07
5.	Grasim, HDFC, HDFC Bank, ICICI Bank, SBI	Diversified, Finance, Banks	1.01+2.94+3.15+4.05+4.98=16.13
6.	Hero Honda, M&M, Maruti Suzuki, Tata Motors	Auto	1.5+1.14+1.57+1.62=5.83
7.	HUL, ITC, Sun Pharma	Personal Care, Cigarettes	1.93+3.78+1.34=7.05

Table 1: Sectors with approximate weights in the SENSEX

The scrips of relevance can be generated from the database by finding the frequent itemsets and then discovering the rules for all itemsets above some minimum support threshold. We have used Dynamic Itemset Counting (DIC)[1] for generating the frequent itemsets.

C. Generation of Association Rules

The association rules between scrips are positively or negatively correlated. These rules recommend to buy stock2 if stock1 is bought, if stock1 and stock2 exhibit positive correlation. If a negative correlation exists between them a rise in stock1 can trigger a sell stock2.

The same rules can be applied for intra-sector and inter-sector recommendations. In each sector, we have companies operating in similar areas: for example in cement, construction, Engineering and steel. They may operate in the same market and a rise in the market share of one may adversely affect the market share of the other. In other circumstances some scrips may show a positive correlation and may rise together as a result of some news positively affecting the entire sector. Similar would be the case for inter-sector association rules.

The association rules so generated have been mined in different support and confidence frameworks.

Examples of Association rules generated are:

For support greater than 20% and confidence greater than 90%:

92.592590 : incr HDFC ^ incr Rel Infra ^ incr Rel Comm ----> incr ICICI Bank

92.592590:decr BHEL ^ decr HDFCBank---->decr ICICI Bank
93.589745:decr power,metals ^ decr cement, engineering, construction,steel---->decr diversified , Finance, Banks

For support greater than 30% and confidence greater than 60%:
Intra-sector rules:
89.705879 : decr HDFC Bank ^ decr SBI ----> decr ICICI Bank
64.000000 : decr Rel Infra ----> incr Hindalco

Inter-sector rules:
100.000000 : decr power,metals ----> decr cement, engineering, construction, steel
100.000000 : incr Oil ^ incr power,metals----> incr cement,engineering, construction, steel

87.012985 : decr Oil ^ decr Telecom,computers ----> decr Cement, engineering, construction, steel

86.826347 :decr Telecom,computers ^ decr diversified,Finance,Banks---->decr cement,engineering, construction, steel

These rules are generated which are transformed into buy or sell recommendations.

D. Generation of rules with time lags

We define a lag as that time after which we calculate the percentage of rise or fall for a particular scrip. This time can be the number of trading days, weeks or months. In our case we represent lag as the number of trading days. There are certain patterns which may not be detected in transactions created on the basis of closing prices of each consecutive day. These patterns are such that they are observed after a particular time lag. Here lag refers to the number of days that have elapsed between the two closing prices. The rules that we have generated above are lag=1 rules. Rules which are observed with a lag greater than 1, like a steady rise/fall, are missed out in these set of rules.

For example a steadily rising stock may rise 0.1-0.3% everyday and may show an increase of 1% after five days. Hence if we calculate the percentage rise on every fifth day this stock will be included in the dataset.

For example two stocks can follow a pattern such that if stock1 rises on day1 then stock 2 follows stock1 on day3 with support and confidence greater than the threshold. So we say that stock2 follows stock1 with lag=3.

To find these rules, frequent itemsets on the different lag datasets are found. Rules are generated on the individual frequent itemsets and only the strongest rules are chosen. The days on which the strongest rules occur gives the time lag for that particular rule. That is we have rules of the form:

If time lag =3, decr SBI --- > incr ACC
If time lag =4, incr HDFC^incr Rel Infra ---- >decr BHEL

These rules will not be observed otherwise while considering the transactions on a day to day basis.

IV. FUZZY ASSOCIATION RULE GENERATION

Fuzzy Association Rule Mining (FARM) is intended to address the crisp boundary problem encountered in traditional ARM.

Due to the crisp boundaries, some items on the boundary are missed out though they contribute significantly to relevant rules. FARM takes care of these items and reflects their contribution in the rules generated.

A. Creation of the fuzzy database

In the earlier creation of the database a crisp boundary dictated the inclusion of an item in a transaction i.e if a scrip rose/fell by a particular amount it qualified to be included as an item in a transaction. But in the fuzzy database each item consists of a scrip with its membership value. The membership value lies between 0 and 1. A scrip is included as an item if it has risen/fallen by a value between a range of values. So an item in a fuzzy database is of the form <N,M>, where N is the scrip id and M is the membership function.

There can be various membership functions. For example, let us consider the following fuzzy membership function for inclusion in the transaction database we considered earlier:

M = 1.0	for rise/fall >=2%
= 0.8	for 1.8 <= rise/fall <2%
= 0.6	for 1.6 <=rise/fall <1.8%
= 0.5	for 1.4 <=rise/fall <1.6%
=0.2	for 1.2 <=rise/fall <1.4%
=0.1	for 1.0<= rise/fall <1.2%
=0.0	for rise/fall <1.0%

So each item in the fuzzy database is of the form <a, 0.5>, <b, 1.0> etc. where a, b correspond to the scrips and 0.5, 1.0 are their membership values.

B. SUPPORT CALCULATION:

The support for a 1-itemset is simply the sum of the membership degree values divided by the number of records in the database. The support for an n-itemset, for each record containing the itemset, is the sum of the products of the membership degree values in each record. Thus for example if we have database records of the form:

<c, 1.0>
<a,0.5> <b,0.5>
<a,0.5> <c, 0.5>
<a,0.5> <b, 0.5>

The calculated support values will be:

{a} = 0.375
{c} = 0.375
{a c} = 0.0625
{b}=0.25
{a b}=0.125

Therefore after the calculation of support for each item, we can calculate the association rules by the earlier method itself.

The objective of fuzzy sets is that it discovers many hidden rules in transactions because fundamentally strong scrips show a gradual rise which is not captured in the crisp boundary support calculation as opposed t speculative scrips.

C. FUZZY TIME LAGGED ASSOCIATION RULES

In crisp time-lagged association rules, the problem of some rules being missed out remains. This problem to some extent is minimized with the help of fuzzy time-lagged association rules. Here the time-lagged dataset is calculated in a similar manner as for the crisp dataset but the inclusion in the dataset is defined by the fuzzy function as also the support calculation. These give the additional rules not captured earlier.

V. WEIGHTED FUZZY ASSOCIATION RULE GENERATION

The large number of frequent itemsets and association rules generated can be pruned using

We have used weighted fuzzy association rule mining for assigning the weights to the 30 SENSEX scrips. These weights are assigned to determine the value of the SENSEX with rise/fall in particular scrips.

These weighted fuzzy datasets help in finding inter-transaction association rules, inter-transaction association rules and also for predicting the value of the SENSEX as some scrips/sectors are more heavily weighted in the SENSEX than the others.

The fuzzy dataset created, as explained earlier, is associated with the weighting file.

A. SUPPORT CALCULATION

For single itemsets, the support is the sum of the product calculation for each weighting/fuzzy membership pair. For 2-itemsets and larger the support is the sum of the products of all the weightings and fuzzy membership calculations.

If data is as follows:

<c, 1.0>
<a,0.25> <b,0.5>
<a,0.5> <c, 0.75>
<a,0.5> <b, 0.25> and the associated weighting file is:
0.2
1.0
0.3

The support calculations would be as follows:

{a} = ((0.25*0.2) +(0.5*0.2) +(0.5*0.2))/4 = 0.0625

Similarly:

{b}=0.1875
{c}=0.13125
{a,b}=((0.25*0.2*0.5*1.0)+(0.5*0.2*0.25*1.0))/4=0.0125
Similarly:
{a,c}=0.005625

B. WEIGHTED FUZZY TIME-LAGGED ASSOCIATION RULE GENERATION

The creation of the dataset is similar. These weighed rules help in associating the rise/fall in the SENSEX with the time lags as these weights represent the approximate weights in the SENSEX.

VI. IMPLEMENTATION OF THE RECOMMENDER SYSTEM

We assume that a portfolio has to be managed with the obvious intention of making a profit. The portfolio can already contain scrips which can be replaced and the portfolio restructured or the portfolio can be created by initializing it with the scrips from different sectors.

Then a time frame for monitoring is fixed. After periodic intervals association rules are generated and loss making stocks can be replaced by corresponding negatively correlated rising stocks of the same amount. The same is applicable to mutual funds.

Also inter-sector rules can help to switch sectors, if some sectors are expected to perform better than the others.

Future direction of work:

The above techniques can be extended to generate predictive rules for intra-day trading and can help in recommending related stocks on an intra-day basis for trading.

Stream mining can be used to generate frequent itemsets and rules to trigger rules on an intra-day basis for recommendation. These rules can be new or those which are already existent. We have used the space saving algorithm using sliding windows to generate these frequent itemsets but have not generated the rules on the stream data.

References:

[1] S.Brin, R., J Ullman, Shalom Tsur,"Dyanamic Itemset Counting and Implicaton Rules for Market Basket data", SIGMOD Record, volume 6, no 2, pages 255-264, June 1997

[2] G. Marketos, K. Pediaditakis, Y. Theodoridis, B. Theodoulidis "Intelligent Stock Market Assistant using Temporal Data Mining", Proc. 10th Panhellenic Conference in Informatics (**PCI'05**), Volos, Greece, November 2005.

[3] Ashish Mangalampalli, Vikram Pudi "Fuzzy Association Rule Minng Algrithm for Fast and Efficient Performance on Very Large Datasets"

[4] P.Velvadivu and Dr. K.Duraisamy,"An Optimized Weighted Association Rule Mining on Dynamic Content", IJCSI International Journal of Computer Science Issues, Vol 7, Issue 2, No.5, March 2010

[5] Hitesh Chhinkaniwala, P.Santhi Thilagam, "InterTARM:FP-Treee based framework for mning Inter-Transaction associaton Rules from stock market data", Internatonal Conference On Computer Science and Information Technology, 2008

[6] Mark Grinblatt, Tobias Moskowitz "Predicting stock Price movements from past returns : The role of Consistency and Tax-Loss Selling", Journal of Financial Economics,71 (2004) 541-579

[7] K.Senthamarai Kannan, P.Sailpathi Sekhar, M.Mohamad Sathik,P.Arumugam, "Financial Stock Market Forecast Using Data Mining Techniques", IMECS 2010, March 17-19, 2010, Hong Kong

[8] Hameed Al-Qaheri, Aboul Ella Hassanien and Ajith Abraham"Discovering Stock Price Prediction Rules using Rough Sets",

[9] Benjamin Wah, Minglun Qian, "Constrained formulations and algorithms for stock-price predictions using recurrent FIR neural networks",Eighteenth National Conference on Artificial Intelligence, Edmonton, Canada, 211-216, 2002

[10] Amit Agarwal, Elad Hazan, Satyen Kale, Robert E. Scapire, "Algorithms for Portfolio Management based on the Newton Method", Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, PA, 2006

[11] Hazan . E, Kalai. A, Kale S**,** Agrawal A, "Logarithmic regret algorithms for online convex optimization", 19th Annual Conference on Learning Theory (COLT), 2006