Applied Soft Computing 13 (2013) 1055–1063

Contents lists available at SciVerse ScienceDirect

Applied Soft Computing

journal homepage: www.elsevier.com/locate/asoc

A stock market portfolio recommender system based on association rule mining

Preeti Paranjape-Voditel [a,*], Umesh Deshpande [b]

[a] Ramdeobaba College of Engineering and Management, Katol Road, Nagpur, Maharashtra, India
[b] Visvesvaraya National Institute of Technology (VNIT), Nagpur, Maharashtra, India

ARTICLE INFO

Article history:
Received 22 September 2011
Received in revised form 10 March 2012
Accepted 22 September 2012
Available online 3 October 2012

Keywords:
Association rule mining (ARM)
Portfolio recommender systems
Fuzzy logic

ABSTRACT

We propose a stock market portfolio recommender system based on association rule mining (ARM) that analyzes stock data and suggests a ranked basket of stocks. The objective of this recommender system is to support stock market traders, individual investors and fund managers in their decisions by suggesting investment in a group of equity stocks when strong evidence of possible profit from these transactions is available.

Our system is different compared to existing systems because it finds the correlation between stocks and recommends a portfolio. Existing techniques recommend buying or selling a single stock and do not recommend a portfolio.

We have used the support confidence framework for generating association rules. The use of traditional ARM is infeasible because the number of association rules is exponential and finding relevant rules from this set is difficult. Therefore ARM techniques have been augmented with domain specific techniques like formation of thematical sectors, use of cross-sector and intra-sector rules to overcome the disadvantages of traditional ARM.

We have implemented novel methods like using fuzzy logic and the concept of time lags to generate datasets from actual data of stock prices.

Thorough experimentation has been performed on a variety of datasets like the BSE-30 sensitive Index, the S&P CNX Nifty or NSE-50, S&P CNX-100 and DOW-30 Industrial Average. We have compared the returns of our recommender system with the returns obtained from the top-5 mutual funds in India. The results of our system have surpassed the results from the mutual funds for all the datasets.

Our approach demonstrates the application of soft computing techniques like ARM and fuzzy classification in the design of recommender systems.

1. Introduction

We have generated a stock market recommender system based on association rule mining (ARM) that recommends a portfolio of stocks. The objective of this recommender system is to support stock market traders, individual investors and fund managers in their decisions by suggesting investment in a group of equity stocks when strong evidence of possible profit from these transactions is available. Our system is different compared to existing systems because it finds the correlation between stocks and recommends a portfolio. The existing techniques based on technical and fundamental analyses recommend buying or selling a single stock based on the price volume patterns on fundamentals of the stock. They do not recommend a portfolio. To the best of our knowledge, ours is the first attempt to use the technique of association rule mining for creating a stock market portfolio recommender system.

ARM has been used with great success in domains such as market basket analysis. However, the use of traditional ARM is infeasible for stock market predictions since the number of association rules generated for the stock data is exponential to the number of itemsets and finding relevant rules from this set is difficult. We have developed an effective domain based pruning technique to reduce the number of rules generated and also to retain their relevance.

Any stock market can be divided into thematic sectors depending on the area of operation of the company to which the stock belongs. The number of sectors depends on the number of stocks, the type of the market and the diversification in the market. The objective of this division is to generate meaningful rules between the sectors and also within each sector. We will call them the cross sector and intra sector rules respectively. These rules are used for the portfolio formation. The division into sectors helps to reduce irrelevant rules.

* Corresponding author. Tel.: +91 9823084251.
E-mail addresses: preetivoditel@gmail.com, voditelps@rknec.edu, preetiparanjape@yahoo.com (P. Paranjape-Voditel), uadeshpande@cse.vnit.ac.in (U. Deshpande).

1568-4946/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.asoc.2012.09.012

In this paper, we first propose a technique based on standard ARM for portfolio creation based on historical data of stock prices. We also propose a technique for rebalancing the portfolio at regular intervals (e.g. after every year) using the cross sector and intra sector rules obtained using ARM, if we observe that some stocks are not performing to their expectations. The results of the techniques were much better than the returns given by a few of the leading mutual funds.

This technique was then improved by preparing a fuzzy dataset where the inclusion of stocks in the dataset from the historical data was based on a membership function. We used fuzzy ARM on this dataset and obtained better results as compared to the earlier technique. After observing the historical data, we also came up with a domain specific technique of time lagging to take into account the impact of slowly rising or falling stocks. The fuzzy ARM based technique for portfolio creation and rebalancing was further augmented by using this technique of time lagging. The use of this augmented technique led to further improvement in the results obtained which surpassed the results given by the leading mutual funds by a huge margin.

Throughout the paper, we have explained the techniques and all the associated terms with a running example on the BSE-30 dataset. BSE-30 (also called the SENSEX), is a free-float market capitalization-weighted stock market index of 30 well-established and financially sound companies listed on the Bombay Stock Exchange. The 30 component companies which are some of the largest and most actively traded stocks are representative of various industrial sectors of the Indian economy.

In order to demonstrate the wide applicability of our technique, we have carried out extensive experimentation of our technique on various datasets in addition to BSE-30. These include Indian datasets of S&P CNX Nifty (which is a well diversified 50 stock index accounting for 22 sectors) and S&P CNX 100 (which is a diversified 100 stock index accounting for 38 sectors). We have also experimented on DOW-30 Industrial Average, which is the second oldest U.S. market index. The Dow Jones Industrial Average is a price-weighted average of 30 blue-chip stocks that are generally the leaders in their industry. It has been a widely followed indicator of the US stock market since October 1, 1928. The results on all the datasets have been excellent and are reported in the paper.

Our approach typically demonstrates the application of soft computing techniques like ARM and fuzzy classification in the design of an efficient recommender system.

The paper is organized as follows. Section 2 gives the background of ARM describing in brief the generation of frequent itemsets and association rules using the support confidence framework. Section 3 describes in detail the creation of a portfolio and its rebalancing. In the same section, we describe the technique using a running example on the BSE-30 dataset. In Section 4, we discuss the fuzzy technique for portfolio generation. Section 5 introduces the concept of time lagged datasets. Section 6 discusses the experimentation and presents a detailed analysis of the results. Section 7 discusses the existing methods for stock market prediction. We conclude the paper with Section 8.

2. Background of ARM

In the problem of ARM, let $I = \{i_1, i_2, \ldots, i_n\}$ be a set of n items and let $D = \{t_1, t_2, \ldots, t_m\}$ be the set of transactions called the database. Each transaction in D has a unique transaction id and contains a subset of the items in I. A rule is defined as an implication of the form $X \Rightarrow Y$ where $X, Y \subseteq I$ and $X \cap Y = \phi$. Here X and Y are called the antecedent and the consequent, respectively.

ARM deals with mining transactions from transactional databases. There are different popular measures of interestingness like support-confidence [22], collective strength [21], leverage [22], conviction, lift [16], etc. We have used the support-confidence framework. In this framework we first mine the database for frequent itemsets, which requires an efficient algorithm as generation of frequent itemsets is the most computation intensive portion of the mining process. Once the frequent itemsets from transactions in a database D have been found by any frequent itemset mining algorithm like Apriori, Dynamic Itemset Counting [16] etc., it is possible to generate strong association rules from them (where *strong* association rules satisfy both minimum support and minimum confidence). This can be done using the following equation for confidence:

$$\text{confidence}\,(A \rightarrow B) = P(B|A) = \frac{\text{support count}\,(A \cup B)}{\text{support count}\,(A)}$$

The conditional probability is expressed in terms of itemset support count, where *support count*($A \cup B$) is the number of transactions containing the itemsets $A \cup B$, and *support count*(A) is the number of transactions containing the itemset A. Based on this equation, association rules can be generated as follows:

1. For each frequent itemset l, generate all nonempty subsets of l.
2. For every nonempty subset s of l, output the rule "$s \rightarrow (l - s)$"

$$\text{if} \frac{\text{support count}(l)}{\text{support count}(s)} \geq \min conf$$

where *min conf* is the minimum confidence threshold. As the rules are generated from frequent itemsets, each one automatically satisfies minimum support. So if we say that the confidence of a rule is 50% it means that for 50% of the transactions containing A and B the rule is correct.

This measure of confidence is used as a parameter to rank rules which are generated on the various datasets. There are many measures other than confidence [23] to measure the interestingness of rules but for our system confidence seemed the most appropriate.

3. The basic technique

The transaction datasets on which ARM is applied is created in the following manner. The daily closing prices of stocks are taken into consideration. Each transaction consists of all those stocks whose closing price on that day has shown a rise or fall of x% or more from the previous day's close. This threshold of x% is decided based on the type of dataset and the fluctuations/volatility in the market. For the BSE-30 dataset this value is taken as 2%. Since each transaction corresponds to the number of stocks that have changed by x% or more on that day, the number of transactions consists of the number of days the stock market has been monitored. One dataset is created for the cross sector data correlating different sectors and intra sector datasets are created for each sector as well.

On these datasets, we have obtained cross sector and intra sector rules using ARM. Cross sector rules are obtained by assigning weights to each sector. The weight of a sector is the sum of the weights of the individual stocks in that sector. To correlate two sectors, we normalize the sectors as follows. Suppose the weight of a sector is w_1% and of another is w_2%, where $w_2 = 2w_1$. Then a 2% rise in the first sector will have a similar effect to a 1% rise in the second sector. For a sector to form a part of a transaction in the cross sector database, we normalize the rise or fall of that sector in proportion to the weight of the highest weighted sector.

Running a frequent itemset generation algorithm on the cross sector dataset gives frequent sectors as items whereas running the algorithm on each sector gives stocks as items. Then we generate rules from these frequent items. Three types of rules are generated – positively correlated rising, positively correlated falling and negatively correlated rules. Positively correlated rising sectors or stocks

are those which rise together, positively correlated falling sectors or stocks are those which fall together and negatively correlated sectors or stocks are such that if one falls the other rises.

We now discuss the method of creating the portfolio from the cross and intra sector rules. The portfolio is created for a minimum period called as the lock-in period, which is defined as the minimum period before which an investor cannot liquidate or sell his portfolio of stocks.

We have assumed that the minimum lock-in period for the portfolio is one year for the BSE dataset. The data set is mined for rules using the support confidence framework. The rules are ranked by confidence. The method of creation of the portfolio is like this:

1. Find frequent sectors by running any frequent itemset generation algorithm like Apriori/Dynamic Itemset Counting on the cross sector file. The rules are generated from these frequent sectors. Here we consider only the positively correlated rising sectors. The antecedent as well as the consequent of the top-k of these rules is taken. The value of k is decided according to the number of antecedent and consequent sectors in the rule. For the BSE dataset we have used $k = 3$.We also make a note of negatively correlated sectors. These are then used for rebalancing.
2. On the top-k sectors so found we run a frequent itemset generation algorithm again on the individual sector datasets. This will generate the frequent stocks in the respective sectors. We calculate the association rules on these frequent stocks and take the top-k' rules in each sector. For the BSE-30 dataset we consider the value of k' as 2. Here too we consider only the positively correlated rising stocks. Negatively correlated stocks are noted since they will be used for replacement when we rebalance.
3. Apart from cross and intra sector rules we also have datasets of all the items. These are required to find the relative order of the stocks which are generated in the sector files above. Relative ordering is required because investment in the portfolio is made in the order in which stocks are recommended. Let us assume that we have the following cross sector rule:

$$s_1, s_2 \rightarrow s_3$$

   Then s_1, s_2 and s_3 represent three sectors on which we would be running a frequent itemset generation algorithm for finding frequent stocks. Now s_1, s_2 and s_3 form a rule with the same confidence. To find out the relative ordering of the stocks this would be generated from each of these sectors we need to go back to the rules generated from the entire database where this ordering would be clear. For investment with a limited amount this order has a special significance as not all stocks would be bought. Only the top few stocks would be bought. When the amount is exhausted we stop.
4. After generating the top-k'' stocks we obtain the amount which is available for Investment in the portfolio. In the Stock Exchange a minimum quantity of stocks called the market lot has to be bought. Initially, we buy the minimum market lot for all the top-k'' stocks if the investment amount permits. If it does not then we buy the minimum market lot of the stocks which consumes the investment. In case we have some amount left after buying the minimum market lot we increase our investment from the first stock of the top-k'' stocks. Since investment is made in the order in which the stocks are ordered, the order decided in step 3 is very important. But if the investment is open ended, i.e. if all recommended stocks are invested in according to their market lots, this order may not be important.

3.1. Evaluating and rebalancing the portfolio

After creation of the portfolio the time for which investment is made is decided. The time of investment taken is variable. Small periods of investment are good for trading and frequent entries and exits from the market do not yield very good returns for long term investments. Therefore some reasonable period, for example one year to two years is taken. This period can be easily extended but a lesser period does not reflect the actual impact of market forces, policies, corrective factors generated by intraday trading etc. on the price of a stock.

If the period of investment is 'n' where 'n' is a multiple of 1 year, then we have evaluation '$n - 1'$ times after every year. Rebalancing is done if required. This is the process of replacing the badly performing stocks. After experimentation we observed that rebalancing done for periods less than a year was counterproductive.

We discuss the method of evaluation and rebalancing in the following steps:

1. Check for the rise/fall in the portfolio after 1 year if the investment is for more than a year.
2. Also check for the rise/fall of the individual stocks.
3. Check for the negatively performing stocks.
4. Generate new rules after addition of this one-year data. Identify negatively correlated cross sector and intra sector rules corresponding to the falling stocks.
5. For each falling stock check if most of the stocks in that sector have fallen. If so we can conclude that the sector faces a slump due to various external factors related to the nature of the sector. In that case a negatively correlated sector has to be chosen from the crosssector rules. On this sector, intra sector rules have to be found for the top-k rules. Here k is not fixed because the stocks obtained from the rules are compared for their price with the stock to be replaced. We replace the fallen stock with a stock whose price is comparable with its fallen value. The value may not match exactly so a proportional number of stocks for the replaced stock are bought.

   Step 5 is the portfolio rebalancing step. From our results on various datasets the actual value with rebalancing was better than one without rebalancing for most of the cases. Rebalancing is done only if stocks have fallen. In cases where all stocks have shown a rise, the composition of the portfolio is not altered.

   At the end of the period of investment one can calculate the returns. This is done by a simple measure called as return on investment (ROI) which calculates the percentage change (increase/decrease) in investment with respect to the initial investment. To calculate ROI, the benefit (return) on an investment is divided by the cost of the investment; the result is expressed as a percentage or a ratio.
6. The return on investment formula:

$$\text{ROI} = \frac{(\text{value of investment after lockin period} - \text{cost of investment})}{\text{cost of investment}}$$

We will define and use two more terms as performance metrics in relation to the quality of recommendation of stocks in the portfolio in terms of the profit namely precision and rebalancing precision.

$$\text{Precision} = \frac{\text{correctly recommended stocks}}{\text{total recommended stocks}}$$

$$\text{Rebalancing precision} = \frac{\text{correctly recommended stocks after rebalancing}}{\text{total recomended stocks}}$$

Here correctly recommended stocks imply that the invested stocks have risen during the period under consideration yielding a profit without rebalancing.

Table 1
Sectors with approximate weights in the SENSEX.

Sector no.	Stocks in the sector	Sector	Approximate weights as per SENSEX
1.	ACC, BHEL, DLF, Jaiprakash Ind, L&T, Tata Steel	Cement, Engineering, Construction, Steel	14.68
2.	ONGC, Reliance	Oil	21.26
3.	Hindalco, NTPC, Rel.Infra, Sterlite, Tata Power	Power, Metals	6.75
4.	Bharti Airtel, Infosys, Rel.Comm, TCS, Wipro	Telecom, Computers	22.07
5.	Grasim, HDFC, HDFC Bank, ICICI Bank, SBI	Diversified, Finance, Banks	16.13
6.	Hero Honda, M&M, Maruti Suzuki, Tata Motors	Auto	5.83
7.	HUL, ITC, Sun Pharma	Personal Care, Cigarettes	7.05

In calculating rebalancing precision we find the stocks which have made a total profit after rebalancing taking into account the initial investment and the final price. The choice of a performance measure depends upon the application domain. For recommendation tasks there are other measures such as accuracy and recall but precision is perhaps the most significant here because higher quality recommendations are more important than recommending a large number of items.

3.2. Portfolio formation on BSE-30

For the technique of portfolio formation, evaluation and rebalancing described in Section 3.1, we take a running example of the BSE-30 stocks [14]. For these stocks, we have formed seven sectors representing stocks from different industries. Each of these stocks has a certain weight in the SENSEX, leading to each sector having a particular weight in the SENSEX. We have formed the sectors as shown in Table 1.

For these sectors we generate the cross sector and the intra sector datasets. We also generate a dataset of all the stocks. A representative subset of the database of all stocks is shown in Fig. 1.

Fig. 1 shows stocks from the BSE-30 which have risen or fallen by an amount of 2% or more. Stocks suffixed with "ˇ" indicatestocks which have fallen and those without have risen.

Steps of portfolio formation applied on BSE-30:

1. Let us consider the rules generated on the cross sector file of the BSE-30 dataset:
   The top-3 positively correlated cross-sector rules are:

   Telecom & Computers, Diversified, Finance&Banks → Auto conf = 89.01%
   Diversified, Finance&Banks → Auto conf = 88.90%
   Power&Metals → Telecom&Computers conf = 86.97%

   These three rules give the top four sectors as Power & Metals, Telecom & Computers, Diversified, Finance & Banks, Auto. We will consider these sectors for the formation of the portfolio from the BSE-30 Index.
2. The top-4 sectors obtained from the top-3 cross sector rules are 3, 4, 5 and 6.On each of these sectors we find intra-sector rules between the different stocks.
   Sector 3:

   Hindalco, Rel.Infra → Tata Motors conf = 78.64%
   Tata Motors, Hindalco → Rel. Infra conf = 78.31%

   Sector 4:

   Infosys, Rel.Comm → Wipro conf = 80.45%
   Infosys → Wipro conf = 78.75%

   Sector 5:

   HDFC, HDFC Bank → ICICI Bank conf = 75.2%
   HDFC Bank, ICICI Bank → SBI conf = 74.1%

   Sector 6:

   Hero Honda, M&M → Maruti Suzuki conf = 85.64%
   Hero Honda, Tata Motors → M&M conf = 80.31%

   Here from each of the sectors we have taken the top-2 rules ranked by their confidence.
3. From step 2 above, the intra sector rules between the stocks have been generated. The method to order them as discussed in step 3 of Section 3.1 is shown on the dataset of BSE-30.
   For example if we have the top-3 rules in the cross-sector rules as:

   Telecom&Computers, Diversified, Finance&Banks → Auto conf = 89.01%
   Diversified, Finance&Banks → Auto conf = 88.90%
   Power&Metals → Telecom&Computers conf = 86.97%

   If we take the first rule, the three sectors are taken on which we run a frequent itemset generation algorithm for generation of intra-sector rules. But for ordering these rules we require the rules from the entire database. The entire database does not contain sectors.
   So the relative order of the stocks belonging to the same rule can be found out. After finding this relative order we order the rules and then take the antecedent and consequent of the top 2 rules from each sector for generating the portfolio.
   For our example, after ordering of the rules we have:

   Infosys, Rel. Comm → Wipro
   Infosys → Wipro

   HDFC, HDFC Bank → ICICI Bank
   HDFC Bank, ICICI Bank → SBI
   Hero Honda, M&M → Maruti Suzuki
   Hero Honda, Tata Motors → M&M

   Hindalco, Rel. Infra → Tata Power
   Tata Power, Hindalco → Rel. Infra
4. Therefore the order of stocks for investment is:
   Infosys, Rel.Com, Wipro, HDFC, HDFC Bank, ICICI Bank, SBI, Hero Honda, M&M, Maruti Suzuki, Tata Motors, Hindalco, Rel.Infra, Tata power

ACC˘ Bharti Airtel˘ dlf˘ grasim˘ HDFC˘ HDFC Bank˘ Hero Honda˘ Hindalco˘ HUL ICICI Bank SBI˘
Infosys L&T Rel Infra˘ ONGC˘ Tata Power Tata Steel˘
ACC Bharti Airtel˘ NTPC˘ ITC TCS

Fig. 1. Dataset of all stocks.

The portfolio generated is as shown in Table 2.

3.3. Evaluating and rebalancing the portfolio from BSE-30

We have taken the minimum lock-in period to be one year. We evaluate the portfolio after one year. In the present example we take the period of investment as two years. After a year we rebalance the portfolio by replacing the stocks which have shown negative growth by a method described below. We have assumed that the investor is interested in long term investment and hence have chosen the period of two years with rebalancing after one year.

1. In our example of the portfolio on BSE-30, two stocks have shown negative growth namely Reliance Infra and Reliance Communications or Rel.Com.
2. As stated in step 3, for our portfolio example, Rel.Infra and Rel.Com have to be replaced. First we look for the rise/fall of other stocks in the respective sectors. Rel.Infra belongs to the Power & Metals sector. Since the stocks in this sector have not all fallen there is no slump in the sector. Hence we try to find a negatively correlated stock from the same sector. Hindalco is the only negatively correlated stock with Rel.Infra but due to the price difference between the fallen price of Rel.Infra and the stock price of Hindalco, we buy 528.58number of shares of Hindalco as shown in Table 2.
3. For Rel.Com from the Telecom & Computers sector, the sector has not shown a slump as all other stocks from this sector have risen positively. So intra-sector rules have to be applied to replace it with a stock of the same sector. For our example Rel.Com is negatively correlated to TCS. Hence with additional investment Rel.Com is replaced by TCS.
4. For our BSE-30 portfolio this value is:

$$\text{ROI(without replacment)} = \frac{(1,673,445 - 1,135,130)}{1,135,130} = 47.42\%$$

$$\text{ROI(with replacement)} = \frac{1,732,459.82 - 1,135,130}{1,135,130} = 52.60\%$$

These are the compounded returns for a period of two years. We annualize this return.

The annualized returns work out to 21.24% and 23.69% respectively.

5. For the portfolio the values of precision and rebalancingprecision are:

$$\text{Precision} = \frac{\text{correctly recommended stocks}}{\text{total recommended stocks}} = \frac{12}{14} = 85.71\%$$

$$\text{Rebalancing precision} = \frac{14}{14} = 100.00\%$$

For our example out of the 14 stocks in the portfolio 12 were correctly recommended, 2 were rebalanced and after rebalancing all 14 showed positive results.

We observe that rebalancing precision has been excellent for the portfolio on BSE-30.

4. The fuzzy technique for portfolio creation

The datasets created earlier involved crisp boundaries wherein a stock was included in the dataset only if it rose or fell by a fixed threshold, for example 2% for the BSE-30 dataset.

The results from this dataset were very good but we realized that they could be still better if we do not have crisp boundaries for inclusion of a stock in a dataset.

In the earlier creation of the datasets a crisp boundary dictated the inclusion of an item in a transaction, i.e. if a stock rose or 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 stock with its membership value. The membership value lies between 0 and 1. A stock is included as an item if it has risen or 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 stock and M is its membership function. This function eliminates the possibility of rejection of an item if its rise or fall falls on the crisp boundary. There are various categories of membership functions. We have used the following fuzzy

Table 2
Returns generated on the portfolio on BSE-30 without and with rebalancing.

Stock Name (1/6/2010)	Price on 1/6/2009	Value	Price on 1/6/2010	Price on 1/6/2011	Rebalancing
Infosys	1167.65	116,765	2625.25	2812.00	
Rel.Com.	248.85	24,885	139.25	93.70(w/o) (1170.70 × 18.84 with)	TCS (738.80) × 18.84
Wipro	145.64	14,564	658.10	448.10	
HDFC	459.70	45,970	2705.90	688.20	
HDFC Bank	1433.00	143,300	1857.55	2389.30	
ICICI Bank	723.20	72,320	838.35	1085.05	
SBI	1879.80	187,980	2210.15	2329.65	
Hero Honda	1366.69	136,669	1917.90	1861.25	
M&M	355.23	35,523	563.90	675.65	
Maruti Suzuki	1040.30	104,030	1259.20	1248.90	
Tata Motors	237.95	33,795	725.65	1079.45	
Hindalco	87.75	8775	146.11	197.30	
Rel.Infra	1300.84	130,084	1042.90		579.60 (w/o)
Hindalco	528.58	*146.11	197.30 × 528.58 (with)		
Tata Power	804.70	80,470	1259.75	1246.30	

Total value (1/6/2009) 1,135,130.00, total value (1/6/2011) (without rebalancing), 1,673,445.00, total value (1/6/2011) (with rebalancing) 1,732,459.82, ROI (without rebalancing) 47.42%, ROI (with rebalancing) 52.6%, annualized ROI (w/o rebalancing) 21.24%, annualized ROI (with rebalancing) 23.69%, stocks replaced by rebalancing on 1/6/2010: Rel.Infra by 528.58 × Hindalco Rel.Com by18.84 × TCS.

<1,0.8><6,1.0><7,1.0><9,1.0><10,1.0><12,1.0><15,1.0><19,1.0><21,0.2><23,1.0><26,1.0><27,1.0><29,1.0><30,1.0><3ˇ,0.8><11ˇ,1.0>
<16ˇ,1.0><22ˇ,0.7><25ˇ,1.0>
<1,1.0><3,0.8><6,1.0><9,1.0><10,1.0><11,1.0><12,1.0><15,0.2><19,0.4><21,1.0><22,0.7><23,0.4><25,0.2><26,1.0><29,1.0><7ˇ,0.2>
<8ˇ,1.0><13ˇ,1.0><16ˇ,0.4><27ˇ,1.0><30ˇ,1.0>

Snapshot of standard transactions:
6 7 8 9 10 12 15 19 23 26 27 29 30 11ˇ 16ˇ25ˇ
1 6 9 10 11 12 21 26 29 8ˇ 13ˇ 27ˇ 30ˇ

1-ACC,2-bharti,3-BHEL,4-DLF,5-Grasim,6-HDFC,7-HDFCBank,8-HeroHonda,9-Hindalco,10-HUL,11-ICICIBank,12-Infosys,13-ITC,14-JP,15-L&T,16-M&M,17-Maruti,18-NTPC,19-ONGC,20-Rel.Com,21-Rel.Infra,22-Rel,23-SBI,24-Sterlite,25-tatamotor,26-tatapower,27-tatasteel,28-TCS,29-WIPRO,30-Cipla

Fig. 2. Comparison of the standard and fuzzy database for the same period.

membership function for inclusion in the transaction database we considered earlier:

$$\begin{aligned} M &= 1.0 \text{ for rise/fall} \geq 2\% \\ &= 0.8 \text{ for } 1.8 \leq \text{rise/fall} < 2\% \\ &= 0.6 \text{ for } 1.6 \leq \text{rise/fall} < 1.8\% \\ &= 0.5 \text{ for } 1.4 \leq \text{rise/fall} < 1.6\% \\ &= 0.2 \text{ for } 1.2 \leq \text{rise/fall} < 1.4\% \\ &= 0.1 \text{ for } 1.0 \leq \text{rise/fall} < 1.2\% \\ &= 0.0 \text{ for rise/fall} < 1.0 \end{aligned}$$

This fuzzy function is dependent on the dataset. It can be modified by the user.

In the above fuzzy membership function a rise or fall below 1% is not included in the dataset. The values in this fuzzy function can be changed depending upon the characteristics of the database. Contrary to this, the standard dataset will consist of only those stocks whose membership function is 1.

The effect of fuzzifying the crisp function can be seen in the results obtained from fuzzy portfolios. We compare the effect of the fuzzy function on inclusion of items in Fig. 2. For the sake of convenience we include ids instead of stock names in the transactions and the corresponding stock names are given in the adjoining table in Fig. 2.

Note the non-inclusion of 1, 3ˇ, 22ˇ in the first transaction of the standard dataset and of 3, 15, 19, 22, 23, 25, 7ˇ, 16ˇ in the second transaction.

4.1. Support calculation of itemsets in a fuzzy database

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⟩⟨b, 0.5⟩⟨a, 0.5⟩⟨c, 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

The portfolio formation steps as discussed for the standard dataset are similar for the fuzzy dataset hence we directly discuss the returns from the fuzzy portfolio in Table 3.

The stocks in the fuzzy portfolio for BSE-30 were Hindalco, NTPC, Tata Power, Bharti Airtel, Infosys, Rel.Com, Wipro, Hero Honda, Maruti Suzuki, M&M, Tata Motors, HUL, ITC and Cipla. Here the rebalancing is done on three stocks namely NTPC, Bharti Airtel and Rel.Com. For NTPC cross sector replacement is done whereas for Bharti Airtel and Rel.Com intra sector replacement is done. The precision is 78.57%and the rebalancing precision is 100%. The overall returns as compared to standard datasets have improved phenomenally.

5. Time-lagged datasets

The fuzzy datasets were an improvement over the standard datasets in terms of inclusion of relevant stocks and also in the results.

There may be some fundamentally, steadily rising stocks or some stocks whose fall is gradual. This may not be captured by data obtained on a day-to-day basis and also the fuzzy dataset.

Hence we have defined time lagged datasets. We define a lag as that time after which we calculate the percentage of rise or fall for a particular stock. This time can be the number of trading days, weeks or months. For our example on BSE-30 we represent lag as the number of trading days. Thus a lag dataset is one where prices are observed at intervals. That is a lag = 1 dataset will mean prices are observed every day, lag = 2 dataset implies prices are monitored with a lag of 2 days, i.e. on the 1–3–5–7, etc. days. Likewise, we have generated datasets till lag = 7. There are certainpatterns 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. For example a steadily rising stock may rise 0.1–0.5% everyday and may show an increase of 1% after three days. Hence if we calculate the percentage rise on every fourth day or after lag = 3, this stock will be included in the dataset. Since fuzzy datasets capture rises/falls between 1% and 2% and above this stock would not be present in the fuzzy dataset, too.

Timelagging can be obtained on standard datasets as well as on fuzzy datasets. Since the fuzzy technique was an improvement over the standard technique we apply time lagging over the fuzzy dataset.

The effects of observing these time lags are evident from the returns obtained from the portfolios.

Table 3
Returns generated on the fuzzy portfolio with and without rebalancing.

ROI (without rebalancing)	65.24%	Annualized ROI (w/o rebalancing)	28.45%
ROI (with rebalancing)	67.88%	Annualized ROI (with rebalancing)	29.61%
Precision	78.57%	Rebalancing precision	100%

Table 4
Returns on BSE-30on the standard and fuzzy time-lagged portfolios without and with rebalancing.

S. No.	Value of lag	% ROI (2 years) without rebalancing		% ROI (2 years) with rebalancing	
		Standard datasets	Fuzzy datasets	Standard datasets	Fuzzy datasets
1.	Lag = 1	47.42	65.24	52.60	67.88
2.	Lag = 2	43.90	56.84	50.00	61.33
3.	Lag = 3	21.00	20.64	25.00	25.64
4.	Lag = 4	45.60	21.00	21.00	25.00
5.	Lag = 5	44.46	75.18	52.20	83.37
6.	Lag = 6	35.60	35.60	20.80	21.00
7.	Lag = 7	27.38	35.60	26.00	22.00

5.1. Fuzzy time lagged datasets

These are fuzzy datasets observed after a time lag. The creation of these datasets is similar to fuzzy datasets combined with time-lagged datasets.

From the time-lagged datasets, only one value may show excellent returns compared to the other lag values. This value is useful only if it surpasses the returns from the lag = 1 dataset. We have taken lag values from 1 to 7 for different datasets but we have generally observed that lag = 5 has given us very good results. This lag corresponds to all changes occurring after a week. We show the effect of time lagging on the returns in both the standard and fuzzy datasets in Table 4 for BSE-30. We would like to state that out of the various datasets that we experimented, time lagging till lag = 7 was observed.

From Table 4 we observe that in the results, the technique of finding timelagged fuzzy portfolios yielded maximum returns for lag = 5. We narrowed down on fuzzy timelagged portfolio management for our recommender system.

6. Analysis of results

The portfolio management recommender system discussed in the earlier sections is independent of the number of stocks or sectors of investment. In other words it is independent of data.

For analysis of the recommender system we have chosen the following broad based indices:

The BSE SENSEX, The S&P CNX Nifty or NSE-50, S&P CNX 100 and DOW-30 Industrial Average.

In Sections 3.2 and 3.3, we have shown portfolio recommendations for portfolios derived from the BSE-30 stocks for different periods of time and for different datasets.

We have divided each of the stocks in the indices into sectors and followed the same procedure for portfolio recommendation. We discuss the returns of our system on the indices of the Indian stock market in Table 5.

Since growth mutual funds [20] can be compared our portfolio recommender system portfolios we compare the returns from the top-5 mutual funds in India to those obtained by our system on various indices.

In Table 5 we observe that on NSE-50 and CNX-100 the results have been extremely good. We also see that the system performs well irrespective of data. The results on BSE-30 have also been very encouraging as they have surpassed the results from the top-5 mutual funds.

We discuss the results from NSE-50 which were phenomenal. There was no need to rebalance as all stocks had shown a rise and there was no need for replacement. So the precision was 100%.

We discuss the sectorization of DOW-30 in Table 6.

We would like to state here that when we compared the US markets to the Indian markets the fluctuations or the volatility in the Indian markets is much more. We had formed the fuzzy dataset with a modified fuzzy membership function ranging from 0.5 to 1.5 (Table 6).

For the portfolio on DOW-30 the precision was 100% and there was no need for rebalancing. We see that though the index DOW-30 fell by 6.26% during that period, the returns on our portfolio were 19.62% (Tables 7 and 8).

7. Related work

7.1. Existing methods for stock market prediction

A lot of techniques are existent for individual stock prediction. Traditional techniques such as fundamental and technical analysis provide investors with some tools for managing their stocks and predicting their prices. Technical analysis deals with price volume patterns for individual stocks. Fundamental analysis is a method where the stock is studied for its fundamentals such as EPS, P/E, P/S, D/E, DIVIDEND YIELD, PRICE/BOOK VALUE, DIVIDEND PAYOUT RATIO, CURRENT RATIO, etc. This is again a method where an individual stock which is chosen and found to be undervalued based on

Table 5
Performance of top-5 mutual funds in India vs. Returns on our system.

Funds performance (1 year)		Funds performance (5 year annualized)	
Fund Name	%	Fund name	%
Reliance gold savings fund – growth	34.97	Reliance gold savings fund – Growth	21.55
ICICI prudential FMCG fund growth	32.06	Reliance Banking Fund Growth plan Bonus	20.49
UTI MNC fund – growth	18.42	Reliance Banking Fund – Growth	20.49
UTI transportation & logistic fund growth	16.78	IDFC Premier Equity Find Plan A - Growth	18.63
Canara Robeco Indigo – growth	16.3	UTI Opportunities Fund Growth	15.64

Funds performance (10 years annualized)		Name of index	Period of investment	Returns annualized
Fund name	%			
Reliance growth fund – growth	33.19	BSE-30	1/6/2009 to 1/6/2011	35.27%
Reliance vision fund – growth	29.21	NSE-50	1/1/2003 to 1/1/2004	121.6%
HDFC top 200 fund – growth	28.14			
HDFC equity fund – growth	27.71	CNX-100	1/6/2006 to 1/6/2007	41.77%
HDFC long term advantage fund growth	27.29			

Table 6
Portfolio formation on NSE-50.

Stock name	Price on 1/1/2003	Value	Price on 1/1/2004	Value
ACC	164.35	16,435.00	258.50	25,850.00
Grasim	246.80	24,680.00	801.30	80,130.00
BPCL	224.50	22,450.00	459.85	45,985.00
GAIL	47.19	4719.00	198.49	17,849.00
Bharti Airtel	11.32	1132.00	55.53	5553.00
Siemens	30.81	3081.00	106.07	10,607.00
Wipro	165.88	16,588.00	175.91	17,591.00
ICICI Bank	140.40	14,040.00	302.75	30,275.00
SBI	268.33	26,833.00	532.46	53,246.00
Hero Honda	261.60	26,160.00	458.90	45,890.00
M&M	28.29	2829.00	98.40	9840.00
Tata Motors	162.55	16,255.00	454.45	45,445.00

Total investment (1.1.2003) 175,202.00, value of investment (1.1.2004) 388,261.00, return on investment 121.6%.

Table 7
Formation of sectors for DOW-30.

Sector No.	Stock in sector	Sector	App. wt. as per Index
1.	Caterpillar, Alcoa, Boeing	Construction & Mining, Metals, Defense, Aerospace	11.70
2.	Exxon Mobil, Dupont, Chevron Corp.	Oil & Gas, Chemicals	14.47
3.	3M, General Electric, United Tech. Corp.	Conglomerate	11.12
4.	Cisco Systems, Verizon, Comm., IBM, AT & T, Hewlett Packard, Microsoft, Intel	Telecom, Computers	21.62
5.	Travelers, JP Morgan Chase, Bank of America, American Express	Finance, Banks, Insurance	9.33
6.	Walt Disney, Wal-Mart, The Home Depot, Proctor & Gamble	Consumer Goods, Broadcasting & Entertainment	12.47
7.	Pfizer, Merck, McDonalds, Kraft, Foods, Coca Cola, Johnson & Johnson	Retail, Pharma, Food, Beverages	19.29

these ratios. Both these analyses deal with individual stocks. Stocks in relation to each other give an insight into the interrelationships that exist between them.

Apart from these techniques, there are many other analysis techniques for stock market prediction. [10] deals with the application of artificial neural networks in stock market prediction. It uses the back propagation neural network (BPNN) algorithm to predit the Stock Exchange of Thailand (SET) index. [9] introduces an information gain technique used in machine learning for data mining to evaluate the predictive relationships of numerous financial and economic variables. Neural network models for level estimation and classification are then examined for their ability to provide an effective forecast of future values. Their results show that the trading strategies guided by the classification models generate higher risk-adjusted profits than the buy-and-hold strategy, as well as those guided by the level-estimation based forecasts of the neural networks and linear regression models. [2] discusses a genetic algorithm optimized decision tree – SVM based hybrid stock market trend prediction system. It compares this hybrid system with ANNs and the naive Bayes theorem. [7] presents the design and performance evaluation of a hybrid decision tree-rough set based system for predicting the next days trend in the Bombay Stock Exchange. [4] And uses the hidden Markov models for prediction. [13] investigates the possibility of discrete stock price prediction using a synthesis of linguistic, financial and statistical techniques to create the Arizona Financial Text System (AZFinText). [15] And uses temporal data mining to generate association rules. It uses event sequences, time series analysis and sequential mining for stock market prediction.

In [11] a Hierarchical agglomerative and Recursive *K*-means clustering method is used to predict the short-term stock price movements after the release of the financial reports. The method consists of three phases. First, each financial report is converted into a feature vector and hierarchical agglomerative clustering method is used to divide the converted feature vectors into clusters. Second, the *K*-means clustering method is used to partition each cluster into sub-clusters so that most feature vectors in each sub cluster belong to the same class. Then, for each sub-cluster, its centroid is chosen as the representative feature vector. Finally, the representative feature vector is used to predict the stock price movements.

Web based approaches are discussed in [5,6]. Ref. [5] discusses the social web mining approach whereas [6] introduces a method of stock market prediction based on sentiments of web users. It scans for financial message boards and extracts sentiments expressed by

Table 8
Portfolio formation on DOW-30.

Stock	Price on 2.1.2000	Value	Price on 2.1.2001	Value
Chevron	83.62	8362.00	85.94	8594.00
Exxon.	78.31	7831.00	89.12	8912.00
3M	94.37	9437.00	119.19	11,919.00
United Technologies	62.50	6250.00	75.25	7525.00
Pfizer	31.87	3187.00	46.13	4613.00
Merck	67.60	6760.00	93.00	9300.00
Johnson& Johnson	92.19	9219.00	102.00	0200.00
Total Investment (2.1.2000)	51,046.00			
Value of Investment (2.1.2001)		1,371,362.00		
Returns on Investment 19.62%				
Value of DOW		11,357.51		10,646.15
Percentage change in DOW	−6.26%			

individual authors. The system then learns the correlation between the sentiments and the stock values.

8. Conclusions

In this paper, we have shown how association rule mining with a support confidence framework can be used to build a stock market portfolio recommender system. Our approach demonstrates the application of soft computing techniques like ARM and fuzzy classification in the design of an efficient recommender system. The results of this system on various datasets like BSE-30, S&P CNX-100, CNX-50 or NSE-50 and DOW-30 and for different time periods have been extremely good and have surpassed the returns generated by top mutual funds. This demonstrates that the proposed technique is generic enough to be applied on any dataset.

The stock market recommender system proposed by us can be extended for use in intraday trading using stream mining. It needs to be explored whether various parameters of the technique like the support, the confidence, the threshold for inclusion of stocks in the datasets and the threshold for the fuzzy function could be decided at runtime using the characteristics of the datasets.

Acknowledgements

The present paper is a revised and extended version of the conference paper: Preeti Paranjape-Voditel, Umesh Deshpande, “An Association Rule Mining based Stock Market Recommender system”, In Proceedings of Emerging Applications of Information Technology, Kolkata, India, 2011, pp. 21–24, IEEE. We are thankful to Meenal Gokhale, Sumeet Sharma and Pratima Khandelwal for their contribution in preparation of the datasets.

References

[2] B. Binoy, V.P. Nair, N.R. Mohandas, Sakthivel, A Genetic algorithm optimized decision tree-SVM based stock market trend prediction system”, International Journal of Computer Science and Engineering 2 (9) (2010) 2981–2988.
[4] B. Nobakht, C.-E. Joseph, B. Loni, Stock market analysis and prediction using hidden markov models. Student Conference on Engg and Systems(SCES), pp 1–4, 16–18 March 2012.
[5] A. Yi, Supervised by Miles Osborne, Stock market prediction based on public attentions: a social web mining approach, Thesis, Master of Science School of Informatics, University of Edinburgh, 2009.
[6] V. Sehgal, C. Song, SOPS: stock prediction using web sentiment, In the Proceedings of Seventh International Conference on Data Mining Workshops, 2007, ICDM Workshops, 2007.
[7] B. Binoy, V.P. Nair, N.R. Mohandas, Sakthivel, A decision tree-rough set hybrid system for stock market trend prediction, International Journal of Computer Applications 6 (September (9)) (2010) (0975–8887).
[9] E. David, S. Thawornwong, The use of data mining and neural networks for forecasting stock market returns, Expert Systems with Applications 29 (2005) 927–940.
[10] S. Soni, Applications of ANNs in stock market prediction: a survey, International Journal of Computer Science & Engineering Technology (IJCSET) 2 (March (3)) (2011).
[11] A.J. Lee, M.-C. Lin, R.-T. Kao, K.-T. Chen, An effective clustering approach to stock market prediction (2010). PACIS 2010 Proceedings. Paper 54. http://aisel.aisnet.org/pacis2010/54
[13] R.P. Schumaker, H. Chen, A discrete stock price prediction engine based on financial news, IEEE Computer 43 (Jan (1)) (2010) 51–56.
[14] http://www.bseindia.com/about/abindices/bse30.asp
[15] G. Marketos, K. Pediaditakis, Y. Theodoridis, B. Theodoulidis, Intelligent stock market assistant using temporal data mining, in: Proc. 10th Panhellenic Conference in Informatics (PCI’05), Volos, Greece, November, 2005.
[16] S. Brin, R.J. Ullman, T. Shalom, Dyanamic itemset counting and implicaton rules for market basket data, SIGMOD Record 6 (June (2)) (1997) 255–264.
[20] http://www.mutualfundsindia.com/topfund_rpt.asp
[21] C.C. Aggarwal, P.S. Yu, A new framework for itemset generation, in: PODS 98, Symposium on Principles of Database Systems, Seattle, WA, USA, 1998, pp. 18–24.
[22] G. Piatetsky-Shapiro, Discovery, analysis, and presentation of strong rules, Knowledge Discovery in Databases (1991) 229–248.
[23] E.R. Omiecinski, Alternative interest measures for mining associations in databases, IEEE Transactions on Knowledge and Data Engineering 15 (Jan/Feb (1)) (2003) 57–69.

Preeti Paranjape-Voditel is currently a PhD candidate at The Department of Computer Science and Engineering, Visvesvarayya National Institute of Technology (VNIT), Nagpur, Maharashtra, India. She received her MTech in Computer Science and Information Technology from the Indian Institute of Technlogy, Kharagpur, West Bengal, India and BE in Electronics from Walchand College of Engineering, Sangli, Maharashtra, India. She is presently working as an Assistant Professor in the Department of Computer Applications, Ramdeobaba College of Engineering and Management, Nagpur, Maharashtra, India. Her research interests include Data Mining, Algorithms and Databases.

Umesh Deshpande received his PhD in Computer Science and Engineering in 2005 from the Indian Institute of Technlogy, Kharagpur, West Bengal, India. He received his Masters from the Indian Institute of Technlogy, Bombay, Maharashtra, India and BE from Visvesvarayya National Institute of Technology (VNIT), Nagpur, Maharashtra, India. He is currently an Associate Professor in the Department of Computer Science and Engineering at Visvesvaraya National Institute of Technology (VNIT), Nagpur, Maharashtra, India. His current research interests include distributed systems, real-time operating systems, multi-agent systems and Data Mining.