Cumhuriyet Üniversitesi Fen Fakültesi
Fen Bilimleri Dergisi (CFD), Cilt:36, No: 4 Özel Sayı (2015)
ISSN: 1300-1949

Cumhuriyet University Faculty of Science
Science Journal (CSJ), Vol. 36, No: 4 Special Issue (2015)
ISSN: 1300-1949

The application of data mining techniques in analysis the stock portfolio in order to identify common patterns in the behavior of shareholders (Case study of selected brokers in Mazandaran province)

Sara Saadati GHASEM SOLTANLO[1,2], Alireza Naser SADRABADI[3,*]

[1]*Department of Management, College of Humanities, Yazd Science and Research Branch, Islamic Azad University, Yazd, Iran*

[2]*Department of Management, Yazd Branch, Islamic Azad University, Yazd, Iran*

[3]*Department of Management, Yazd University, Yazd, Iran*

Received: 22.03.2015; Accepted: 29.05.2015

Abstract. In this study, we examined the analysis of stock portfolio stakeholders in order to identify common patterns in the behavior of shareholders. Information required about the shareholders shopping cart was collected from the selected brokers in Mazandaran province / Sari city. This information includes demographic information, such as gender, occupation and education, along with a basket of shares purchased during 2013. Data were collected for 150 shares that during this period have traded at least 5 times. This study for the first time with the use of information about the stock portfolio shareholders has offered a model to identify the pattern of purchase behavior of shareholders. In the proposed model, at first shareholders, according to the number of shares purchased categorized into three levels of risk appetite (high, medium, low) then, with the help of data mining classification techniques and association rules have been extracted rules that can be used as a model to predict the behavior of shareholders.

Keywords: Data mining, association rules, risk appetite

1. INTODUCTION

Economic growth and development are important phenomena of our time, which is considered among the aspirations of many countries. The realization of this necessary and undeniable fact, in this present era need to appropriate mechanisms. One of the essential mechanisms which have a major role in realizing and achieving economic growth and development is the capital market. In fact, the persistence of capital market activity was to increase countries economic growth. Investing in securities is one of the investments, on the one hand, due to the positive affects that in providing the financing markets suitable for investment and on the other hand, creates the applicants' funds and financial resources have of particular importance. Stock Exchange is the means an official of the capital market in which it is traded, stocks and bonds of companies under specific rules and regulations. (Karim Zadeh, 2004). There have been many attempts to predict the stock market information by using traditional statistical methods, but these methods are not sufficient to analyze this amount of information. Data mining is one of the most important tools of information technology in today's competitive business world. This method is able to detect hidden patterns and predict future trends in behavior in the stock market. (Rasoulian and Fathi Gohardany, 2008). Analyzing the behavior of shareholders from stock database is very difficult and challenging. This study is looking for a model for analyzing the behavior of shareholders with the use of databases. Question about this study, is the purchase of the shareholders during the formation of stock portfolios follows a certain pattern or not? For example, a shareholder who is

* Corresponding author. *Email:* Alireza_naser@yazd.ac.ir

Special Issue: Technological Advances of Engineering Sciences

known with a characteristic has purchased stocks of oil products, coke and nuclear fuel, as well as whether to purchase the shares of chemical products or not? Many of managers are interested, that analyze the behavior (purchases) of its customers. A conventional analysis which is done on the database transaction was to find a collection of items that, most likely are purchased by a customer.

Dependency rules

Extraction of correlation rules or dependence is a data mining operation which deals with the search for relationships between features in the data set. Another name for correlation analysis is the shopping cart analysis. In other words, dependency analysis is the study of attributes or characteristics that are associated with each other and is looking to extract the rules of this specification. This procedure is followed by extraction rules, in order to quantify the relationship between two or more properties. Association rules, defined as, if and then, with two supports and confidence criteria. (Ghazanfari et al., 2013) The general form of association rules is as $A \rightarrow B$, which is indicative of an event at the same time among the items A and B. (A and B are two of the favorite items of data storage). A and B, respectively, is called forward and inferior law. Presented various criteria to assess the accuracy and value of rules based on them, you will have good laws among a wide range of possible rules. The best known and most widely used of these criteria is the two criteria of the degree of support and the degree of confidence. Supported by a set of items, such as A, is the ratio of the number of transactions, including all items in A, to the total number of all transactions. Degree of confidence a rule $A \rightarrow B$ is defined as the ratio of the number of repeat both A and B, to the number of occurrences of A to alone, which is also included the fraction of transactions involving A and B, which are acceptable values for both the above criteria. (Fakhr Ahmad, 2006). In most cases, only the rules interesting and useful to us that include items with high frequencies, not items that are rarely found in the data warehouse. Most methods, they assume that we are also looking for a set of items that occur at least at a reasonable fraction of their transactions. In other words, support them, not less than the minimum criterion of our support. The so-called frequent item sets used for the items with high support.

Research History

Mack et al. 2011 were using the association rules algorithm to predict the behavior of shareholders in a financial firms market in Hong Kong. The results show that the rules have been discovered, not only increase the workflow of a financial company, but also lead to deeper understanding of investment behavior. Thus, a finance company is able to customize the very best products and services to customers and based on the extracted rules.

Song and Su (2011) also use of association rules to predict changes in equity between the Korea Composite Stock Price Index and global stock market indexes. It is expected that the rules established to facilitate the decision on the purchase or sale of shares. This study shows that the use of larger sample sizes raw data, not only better than the results, it can be helpful to find patterns and unexpected rules.

Lee et al. (2008) showed that how to use of these law to discover repeated patterns of investment behavior in the Shanghai stock market and how our algorithm, we apply on these rules for data collection purchase of actual stock. This study will help financial institutions to search for patterns to know how to build a portfolio and how learn more about behavioral finance.

Safer (2003) with the use of data mining should be taken to predict abnormal stock market returns. He used the data mining methodology CRISP and models of neural networks that could provide a good approximation of market inefficiencies.

METHODS

There are various ways to implement and execute a data mining project. One of the methods is very strong is methodology CRISP (Jarola, 2011). In this paper, the proposed model according to CRISP, which consisting of six phases. Each of these phases, it is included following sectors. Need to move back and forth between different phases because the input of each phase is dependent on to the output of the previous phase. Each of the six phases shown in Figure 1. Then, the focus of the review each of the phases of this model. (Rasoulian and Sharayei and Fathi Gohardany, 2008)

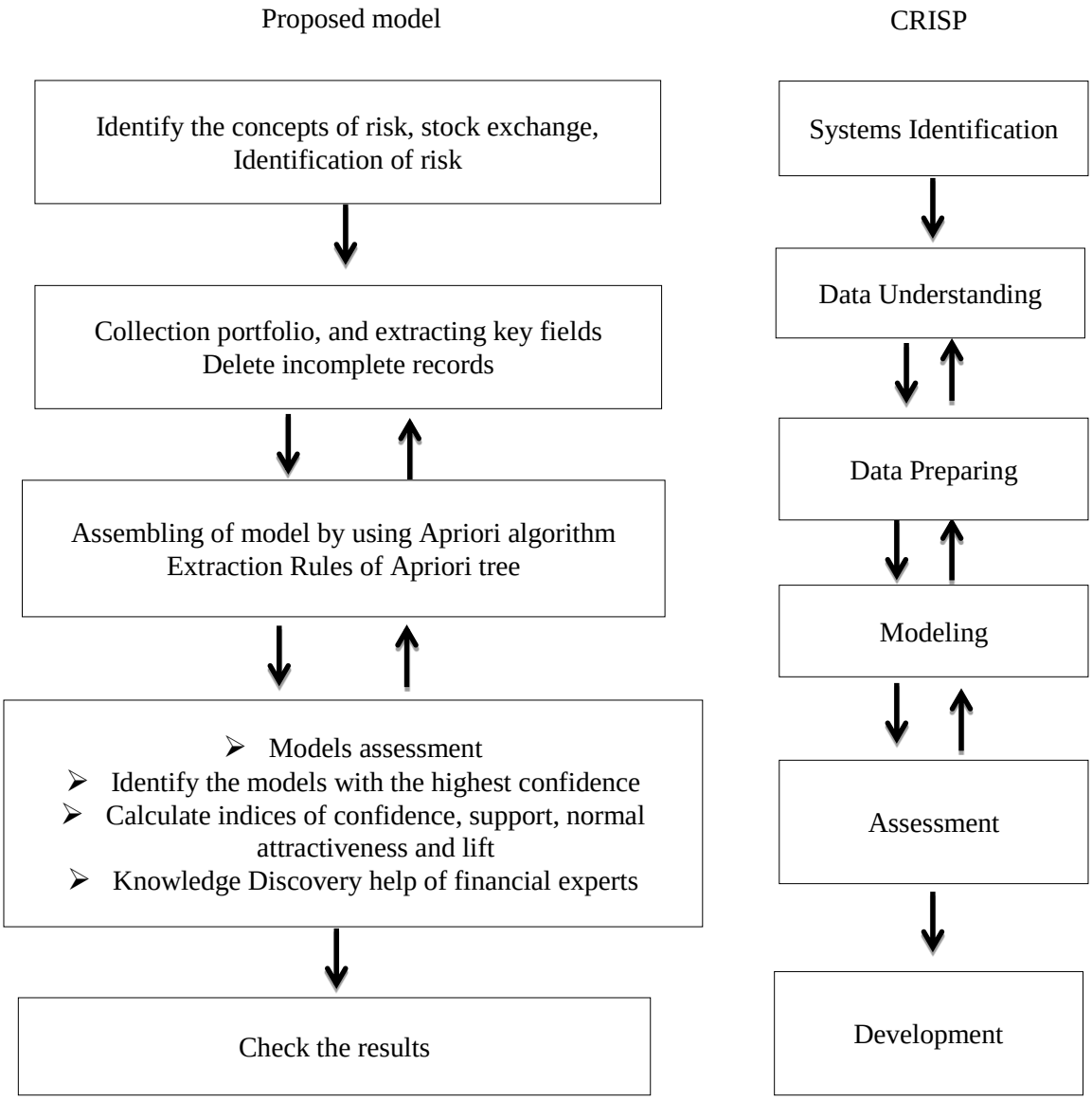

Figure 1. Steps of CRISP methodology and proposed model.

A) Systems Identification

In this phase is focused on understanding the system, then, is to identify and revisited the desired goals and key success factors of the system. In this study, we examined the question is that how

do we pay to identify the model portfolio and whether shareholders with the same risk level, follow the same buying pattern is the most important goal of this research.

B) Identifying data and preparing them

In this step, will be to collect the initial data, describing data inspection and data analysis and data quality validation. The information collected from the three brokers from the center of Mazandaran province / sari. Information portfolio is from 2013. There are about 200 portfolios of shareholders, which after refining shareholders based on the number of transactions over 5 times were selected portfolio of 150 shareholders with privacy stakeholders. Table 1 shows demographic information of collecting samples. The most important step in the investigation (prepares data or pre-processed data) were used to evaluate the shopping cart shareholders. Removed some unnecessary data and poor data.

Table 1. Distribution of frequency percentage variables.

Frequency percentage	Frequency	Sex	Frequency percentage	Frequency	Education	Frequency percentage	Frequency	Job
33.33	50	Female	12.67	19	Diploma	21.33	32	Investor
66.67	100	Male	31.33	47	Associate Degree	46.0	69	Employee
			15.33	23	License	32.67	49	Free
			4.67	61	Master's degree or higher			

C) Modeling

There are many data mining methods for modeling. In this step, by using different techniques of data mining described to the drawing model and optimal pattern. Modeling was performed by using SpssClementin12.0 software. With the increasing diversity of shares purchased decreases the risk of shareholders. (Jafari Samimi and Yahyazadeh and Aminzadeh, 2005). Therefore, in this study, we have used a variety of shares purchased, as a measure to assess the level of risk appetite. Based on this criterion, customers were divided into three groups that following table shows the features and risk appetite of each group. Method of determining the labeling category (level of risk appetite) in the created model is described in Table 2. According to this classification a member of a sample of 150 shareholders, number of 78 people had a high-risk, 43 people had a medium risk and 29 people had a lower risk.

Table 2. The method of people separation based on risk appetite.

Description	Risk
Diversity of shares purchased during a year, less than or equal to 20 shares	High
Diversity of shares purchased during a year, between 20 and 40 shares	Medium
Diversity of shares purchased during a year, more than 40 shares	Low

D) Assessment

Algorithms that are used to get the dependency rules that have the potential to produce a lot of the patterns and rules. In particular, by increasing the number of features (objects) in a data set, it

is possible to produce a large number law, which may necessarily all of these laws, are not attractive for us. Hence, we should be looking at laws that have the most appeal to us. Since the appeal rules depending on factors such as a person use of laws, as well as an area that data collection in this study belong to it, it does not seem to work finding interesting rules. Therefore, it is important that we presented accepted standards to assess the quality of dependency rules. In general, the evaluation criteria can be divided into two categories. The first set of criteria that are often based on human understanding, especially experts in a particular working area. This means that the law can be an interesting one, and for another useless. Determine the categories of standards, fully conforming to the user's taste. Using techniques based on imaging, as well as strategies that are determined to follow the rules of association with terms or restrictions are included in this category. Most of the literature on these criteria, it is believed subjective or mental. The second set of measures, referred to objective criteria which are based on the data in other words, the data are inferred with a clear definition. Values of confidence and support and lift are a bunch of these criteria. Most of the values of these criteria are calculated based on the content of a table with the name of the table is dependency. (Esmaeili, 2013).

a) Support

The ratio of the number of records (transactions) in which the objects B and A, are present both in the total number of records achieved with the support of an important criterion. This numerical value is between 0 and 1, which is as much this amount indicating that the two artifacts, mostly linked together. Using a threshold value, it is weak, it may be, there are laws that support them is less than the threshold, but is valuable legislation.
This is derived from the following equation:

$$(1)\quad \text{Support}(A \rightarrow B) = \frac{\sigma(A \bigcup B)}{N}$$

b) Confidence

Another criteria was to confidence independence rules, which can be obtained from the following equation:

$$(2)\quad \text{Confidence}(A \rightarrow B) = \frac{\text{Sup}(A \bigcup B)}{\text{Sup}(A)}$$

Confidence is a number between 0 to 1, that the more this amount will be added to the quality of legislation. For example, 98% of legal certainty indicates that, in 98% of cases, if it is correct the left law, it will be also true right law. Using this criterion, together with support, is a perfect complement for the evaluation of dependency rules. From disadvantages of this method is that it is possible, there is a high confidence of the law, but not worthy of us. It is even possible, there are laws that from the perspective of qualified persons are considered valuable, but necessarily not have high confidence and support.

c) Lift

Another evaluation criteria dependency rule is a measure of Lift, which is also sometimes known by the name of the Lift. This measure is calculated by the following equation:

$$\text{(3)}\quad \text{Lift}(A \rightarrow B) = \frac{\text{Confidence}(A \rightarrow B)}{\text{Support}(B)}$$

As is clear from the above definition, the amount of lift obtained by dividing the confidence of law on the support of the right law side. In fact, this criterion gets a level of independence between the objects A and B, which this value can be between 0 and infinity. Values close to 1, indicating that they, A and B, are independent of each other and therefore do not show attractive Act. If this criterion is less than 1, indicating that A and B have negative correlation with each other and whatever value of this criterion more than 1, which indicates that A, provides more information about B, which in this case, the attractiveness of the law of BA are estimated to be higher.

E) Development

Build a model, not the end of the study and the purpose of data mining projects is a knowledge discovery and application of knowledge discovered in the future. Knowledge discovery should be organized and also comes in the form usable for others. Extracted patterns help financial institutions and brokers of the stock exchange, as can be use the rules for portfolio investors and gain more information in the field of behavioral finance.

RESULTS

There are various types of association rule algorithms, which could be used to discover patterns in classification. In this study, we used by apriori algorithm. The results of the apriori algorithm on the support of 30% and confidence of 70% are 52 laws to identify and eliminate the weaker laws have been used to lift criterion. However, this amount is closer to one; the law of target is less lift. A set of rules derived in this study, all the laws obtained in the first stage is more lift than 1; therefore, none of the rules will not be deleted at this stage. In the next step is normalized the amount of lift between zero and one. After normalizing of lift rates were excluded from the legislation that had lift less than 0.2. After performing this step obtained 23 final rule which can be seen in Table 3.

Table 3. Dependency rules.

Lift Normal	Lift	Confidence Criteria	Support Criteria	Part of law result	Part of law conditional	Row
00.1	2.52	62.222	30	Mapna Electricity	Pakshoo	1
0.59	1.93	57.778	30	Tamin Pharmaceutical	Mapna Electricity	2
0.58	1.91	52.174	30.667	Tamin Petrochemical, Saderat Bank	Parsian Oil and Gas	3
0.55	1.87	56.14	38	Caspian Sea Shipping	Mapna Electricity	4
0.55	1.87	51.02	32.667	Mapna	Fars & Khuzestan Cement	5
0.53	1.84	55.102	32.667	Mapna	Tamin Pharmaceutica	6
0.43	1.70	50.877	38	Caspian Sea Shipping	Tamin Pharmaceutica	7
0.43	1.69	55.319	31.333	Iran Transfo	Mapna	8
0.41	1.66	54.348	30.667	Tamin Petrochemical, Saderat Bank	Mapna	9
0.39	1.62	61.702	31.333	Iran Transfo	Saderat Bank	10

Lift Normal	Lift	Confidence Criteria	Support Criteria	Part of law result	Part of law conditional	Row
0.38	1.62	50.82	40.667	Bandar Abbas Oil, Tamin Pharmaceutica	Esfahan Oil	11
0.37	1.60	50	30.667	Tamin Petrochemical, Saderat Bank	Iran Transfo	12
0.35	1.57	51.111	30	Mapna Electricity	Mapna	13
0.34	1.56	51.064	31.333	Esfahan Oil	Mapna	14
0.34	1.56	50.82	40.667	Bandar Abbas Oil, Tamin Pharmaceutica	Mapna	15
0.29	1.49	73.469	32.667	Mapna	Bandar Abbas Oil	16
0.28	1.47	72.34	31.333	Iran Transfo	Bandar Abbas Oil	17
0.26	1.44	54.717	35.333	Bahman Group, Tamin Petrochemical	Caspian Sea Shipping	18
0.24	1.41	69.565	30.667	Tamin Petrochemical, Saderat Bank	Bandar Abbas Oil	19
0.24	1.40	53.333	30	Tamin Pharmaceutica	Saderat Bank	20
0.23	1.40	53.061	32.667	Mapna	Saderat Bank	21
0.22	1.38	52.459	40.667	Bandar Abbas Oil, Tamin Pharmaceutica	Saderat Bank	22
0.22	1.38	68.085	31.333	Esfahan Oil	Bandar Abbas Oil	23

According to the rules generated in Table 3, can be done the following interpretations for some of the rules:

Law No. 1: Pakshoo → Electricity Mapna

Law No. 1 shows that people who have purchased the Pakshoo stock have attempted to purchase Mapna Electricity stock. 30% of existing data support the legislation. 45 people from between 150 people have in this pattern of purchase. Take a look at the level of risk appetite of people who have purchased these shares with each other, leading to significant results. 62% of these people are low-risk, and 38% of them are medium risk.

Law No. 8: Mapna → Iran Transfo

Law No. 8 shows that 55.3% of people who have purchased the Mapna shares have also begun to purchase Iran Transfo shares. 33.3% of existing data support this legislation. 45 people from between 150 people have in this pattern of purchase. Take a look at the level of risk appetite of people who have purchased these shares with each other, leading to significant results. 64% of these people are low-risk and 36% of them are medium risk.

Table 4 shows a summary of the results of the analysis of risk levels and rules were extracted. As can be seen in a limited number of rows in the table are present all purchasers with any level of risk. In some laws were the only consideration customers who have a high level of risk and medium and in some laws, there are customers that are low-risk and medium level. In other words, the law is rarely to be found at that level, there is no ability to separate risk clients.

Table 4

law	Low risk	Med risk	High risk
1.00	0/62	0/38	0
2	0/62	0/38/3	0
3	0/64	0/34	0/02
4	0/62	0/35	0/03
5	0/63	/25	0/12
6	0/60	0/28	0/12
7	0/62	0/33	0/05
8	0/062	0/33.3	0/05
9	0/60	0/21.3	0/19
10	0/62	0/18.2	0/20
11	0/62	33/3	0/05
12	0/63	25/6	0/12
13	0/60	0/30	0/10
14	0/60	0/26	0/14
15	0/62	0/10	0/38
16	0/62	0/32	0/06
17	0/62	0/11	0/27
18	0/61	0/33.3	0/06
19	0/63	0/22	0/15
20	0/64	0/29	0/07
21	0/62	0/14	0/24
22	0/63	0/37	0
23	0/60	0/40	0

CONCLUSION

On the question of research, it can be said that the stakeholders in the formation of your shopping cart, follow of pattern of purchase. Rules derived from the association rules describe this pattern. These rules, regardless of the factors influencing on the behavior of shareholders pay to describe the template. Extracted patterns help financial institutions and brokers of the stock exchange and to use these laws for the formation of portfolio investors and gain more information in the field of behavioral finance. It is also expected that the rules established to facilitate the decision on the purchase or sale of shares. Based on the results of section of association rules, we can infer that shareholders with a low level of risk, follow of similar buying patterns and these patterns were observed among all stakeholders with a low level of risk.

REFERENCES

[1] Alizadeh S, Ghazanfari M, Temiorpour B. Data minig and knowledge discovery, publication of iran university of science and technology . 2nd ed. 2011 [In Persian]
[2] Asmaili, M. (2013). Data Mining Azad University of Kashan first .chap [In Persian]
[3] Gafari Smimi Ahmed Mahmoud , Yahya Rahim Amin Zadeh Fr- born Sarynjanblgv- (the relationship between the size of the portfolio and systematic Rsik common shares, the magazine (2005) [In Persian]
[4] Ghazanfari, Mehdi Alizadeh. Temorpor, Babak. (2013). University of Science and Technology, First Edition [In Persian]
[5] Liao s.H.,chen,C.M,W,CH.2008,Mining customer knowledge for product line and brand extension in retailing,Expert systems with Applications,vol.34,no,3,pp.1763_1776
[6] Rasoulian, M., Shrayy, A., Fathi Gvhrdany, MB (2008). "The role of data mining association rules based on research in strategic management, financial, strategic, No. 39 [In Persian]
[7] Sung.H.N.So, Y.S.2011.Forecasting changes in korea compostite stock systems with Applications, vol.38, no, 7, 9046_9049
[8] Sadr, Ahmed Fakhr M., Zolghadr Jahromi M.h., (2006). Rapid method for mining association rules embodied in data collection using logic operations, Twelfth International Conference of Computer Society of Iran, martyr Beheshti University, Department of Electrical Engineering (Computer, Tehran, Iran, 1 to 3 Esfand 85) [In Persian]
[9] Tehran Stock Exchange, the information in the field of capital market in Iran [In Persian]