Prediction of Investment Patterns Using Data Mining Techniques

Sakshi Singh, Harsh Mittal, and Archana Purwar

***Abstract*—Risk taking capability of a person in the financial market is based on many factors including demographic factors like age, education, marital status etc. In this paper in order to analyze the various investment instruments used by the people of different profiles; we have applied fuzzy data mining technique to the demographic factors of a human being. After the division into fuzzy clusters, membership of the person to the clusters is calculated. After the memberships, the derived memberships are used to find the people those have the memberships in the similar range. The result is in the form of investment patterns of the similar people .In further processing the FP growth is applied to find the most recurring patterns. The process sequence for the same has been shown in the paper through the example of five input sets. The main aim of this paper is to apply data mining to find the investment patterns of a person based on their characteristic.**

***Index Terms*—Fuzzy c means, frequent pattern growth, data mining.**

I. Introduction

People of different profiles have different investment strategies as per the humanly attributes. [1], [2] The work in the paper concerns the analysis of the investment patterns which are affected by the attributes of a person it carries in the given population. The classification of the population into groups as per the profile attributes and then analysis of the patterns in investment needs to be done. For any person, to find the investment patterns of the people in the population, the association of the person to the group of similar people need to be found out and then the patterns above a set threshold to be reported.

II. Related Work

Work has been done in the field of risk ability identification of the person. It's more of a subjective matter and depends on the characteristics and circumstances. [3], [4] The factors of inclusions are debated but still questionnaire have been developed with 13 and 5 questions to judge the risk ability and risk tolerance of the person [by John Grable and collaborators]. And they are called the risk assessment instruments.

The allocation task regarding the financial instruments is an elaborate and an exhaustive one which theoretically starts with the utility function and the mapping of the indifference curve. Then the appropriate portfolio selection that leads to efficient indifference curve. Then is the combination between the risk free and the risky class to result in the form of optimum allocation of the financial instruments. The result from CAPM models helps to understand the relation between the risk and the return. [5], [6] of and all the equilibrium between the desired risk and the expected risk and return from the combination of the instruments is matched and results in the portfolio development.

Manuscript received August 8, 2013; revised January 26, 2014.

The authors are with Department of Computer Science /Information Technologg, Jaypee Institute of Information Technology, Noida, India (e-mail: sakshisingh29@gmail.com, harsh5890@gmail.com, archana.purwar@jiit.ac.in).

III. Proposed Work

Data mining is the process of extraction of useful information from datasets. This useful information may include hidden patterns which may help predict future results or a characteristic of a particular person. Data mining includes techniques like Association, Classification and Clustering. The main focus of this paper is on fuzzy clustering algorithm and further extraction of patterns is done by implementing FP growth algorithm. Data mining has been proposed to be used in finance and e-commerce to accomplish new motives [7].

Here rather than extensive work on the risk identification and equilibrium selection between the desired and expected risk and return, data mining techniques are being used to derive the appropriate results. A simple extraction of investment patterns of the like people from the population. The result helps the person to invest the savings in the instruments and how it should be distributed among the same instruments.

Through fuzzy clustering population is divided into groups of similar attributes. [8], [9] Then from the similarities among the attributes taken into consideration similar people are to be found out and patter to be extracted by FP growth. [10] The result is list of people with the like memberships. Few of the attributes taken into consideration are:

- Human factors

Age of the person, Marital Status, Gender, education, family structure

- Income control factors

Occupation, work status, wealth, liquid assets, housing status [11]

Further patterns of the filtered people are analyzed and the most frequent patterns are further filtered in the form of results to the user. This process of extraction of patterns is done through FP growth algorithm and the results are the end results for the user.

IV. Steps Followed

A. Data Collection

The data was taken from the survey of consumer finance.

[11] The survey is taken every 3 years in USA. Raw data of the year 2004, 2007 from the repository of SCF was used. [12] The data collected was the survey of 22,595 people and their attributes like household factors, income and demographic features and also their investments in different markets.

B. Standardization of Data

In this step the standardization of data was done. In standardization, we reduced the data to a [0, 1] range, a method so that the weight of one attribute does not counter the other. A collection of numerical data is standardized by subtracting each value from the mean value and dividing it by its standard deviation. This value obtained is not in [0, 1] range. Subtracting the value from the minimum value and dividing it by the difference of the maximum to minimum value of the attribute gives the desired result. The new values obtained forms the new data, without affecting the hidden patterns inside the dataset. The equations for the standardization used, [13].

$$x_i = ((x_i - \text{mean } X) / sd \quad (1)$$

$$y_i = ((y_i\text{-min}) / (\text{max-min})) \quad (2)$$

In equation 1 x_i is the attribute in the input set, mean$X$ is the mean for the *ith* attribute, *sd* is the standard deviation. In equation 2 y_i is the *ith* input attribute with maximum and minumum for the respective *ith* attributes.

The different attributes of the people were standardized before clustering through the above steps.

C. Fuzzy Clustering

In our work we used fuzzy C means algorithm to get the membership of each tuple in the dataset to the formed clusters. [9]

Certain areas considered while implementation of fuzzy clustering for the model were:

Number of clusters, fuzziness factor, and number of iterations, membership and distance calculation functions.

For the fuzziness factor, due to the absence of experimentation results it was decided to set (fuzziness factor) m=2 is a good choice. [14]

The termination condition for the fuzzy clustering was to fix the number of iterations to be executed, it was analyzed using the data mining tool and was fixed to approximate 25-30 iterations.

For the distance calculation in the case of similarity formulation, Euclidean distance formula was used and for the implementation of the membership function fuzzy c-means was used. [9]

Update Membership:

$$U_{ij}=u_{ij}{}'/_{k=1}{}^{nc}\sum u_{ik}{}' \quad (3)$$

where

$$U_{ij}{}'=(1/(||x_i\text{-}c_j||))/_{j=1}{}^{nc}\sum(1/(||x_i\text{-}c_j||))^{(1/(m\text{-}1))}*U_i \quad (4)$$

$U_{ij}{}'$ is the new membership to the *jth* cluster, x_i is the *ith* attribute and c_j is the respective attribute of the *jth* cluster and m is the fuzziness factor.

Center calculation:

$$C_j=_{i=1}{}^{n}\sum (u_{ij})^m x_i/_{i=1}{}^{n}\sum(u_{ij})^m \quad (5)$$

where C_j center of the cluster, x_i are the ith attribute and u_{ij} is the membership of the tuple to the jth cluster.

D. Extraction of Similar People

As per the membership defined from the training data, every new user was to be assigned the membership to the different clusters. For the membership function, Euclidean distance of the user was calculated from the new centeroids. The distance was calculated by (4). The distance calculated was further used to derive the membership of the new tuple to the different clusters.

The recommendation for the strategies could be done only after the extraction of similar risk group as the user analyzed for its investment strategies. Therefore, the calculated memberships are used and from the clustering results those rows are extracted that have the memberships to the clusters within the threshold limit.

E. Categorization of the Data

Investment of people in different financial instruments in the dataset were recorded in numerical value so to find the hidden investment patterns through FP growth algorithm, there was a need to change the data into categorical form.

The investment strategies of the people in the financial instruments appearing in the dataset are analyzed and has been categorized into different groups based on the criteria of their investment in particular instrument to their financial assets. Values of different financial instruments were divided by the total financial assets of the people and then according to the ratio value, an alphabet is assigned to the financial instrument. This categorical data is then used to identify the patterns. The ratio are equally divided into five categories and respectively represented by the respective starting alphabets.

F. Frequent Pattern Growth

From the extracted group, the patters of investment are extracted that bore the successful results. The criteria for success are the positive returns in the form of unrealized capital gains. The patterns are the static results for the investment to be executed by the users. The investments strategies of the extracted group in different financial instruments are studied and analyzed. FP growth algorithm is used to form the tree of frequently appearing patterns. [10] These patterns extracted are the most frequent ones which are used by the investors.

FP growth algorithm initially calculates the frequency of each item set present in the dataset and then it scans the rows for the formation of a tree. It maintains a header table in which item sets are arranged in ascending order of their frequencies. It displays patterns which have the minimum support fed.

The result set has different investment strategies which are suggested to the user based on FP growth algorithm executed on the tuples which have similar attribute combination leading to similar risk profile. Based on the user financial assets we suggest user optimum amount of money to invest in the instruments.

V. Experimental Results

The given five input sets are the examples tested and the results of the same are stated. (Shown in Table I-Table III).

Assumption: We have taken 5 clusters

All the inputs should be numerals.

Input: For the detailed explanation of the inputs we can be refer. [12]

TABLE I: Inputs

Inputs (16 listed)					
Attribute	***A1***	***A2***	***A3***	***A4***	***A5***
Age	39	44	52	27	40
Edu	12	12	16	14	15
Marital	No	No	Yes	No	Yes
Kids	2	3	3	0	1
Fs	4	4	2	4	4
Oc1	1	1	1	3	1
Oc2	3	2	1	4	2
Income	84216.16	28756.76	63675.68	23621.62	567473.45
Nfa	161300	15000	220700	2800	54300
Assets	174100	16500	335580	3400	342610
Debt	80800	13400	51050	0	890
Nw	93300	3100	284530	3400	6742300
Debt/income	0.123	0.136	0.1896	0	0.012
Fa	11000	1500	114880	600	123800
Bus	0	0	0	1200	80000
Cg	0	0	0	0	1456

**nfa: non financial assets
Fa: financial assets
Fs: family structure
Oc1/Oc2: occupational category 1 and 2
Edu: Education
Nw: Net worth
Bus: Business
Cg: Capital Gain

First of all the input sets were standardized before further processing.

Output: The input sets are executed to find the distances from the centroids of the clusters formed after clustering of the population.

TABLE II: Memberships

Inputs sets		***Clusters***				
	A_i	C_1	C_2	C_3	C_4	C_5
1	A1	0.0355	0.0481	0.8307	0.01813	0.0676
2	A2	0.0452	0.0589	0.0732	0.7516	0.0712
3	A3	0.0568	0.1039	0.0950	0.6199	0.1243
4	A4	0.3172	0.1702	0.2833	0.0669	0.1623
5	A5	0.0936	0.1849	0.14036	0.3919	0.18923

Output: After the FP growth applied to the investment patterns of the people extracted with similar attributes

TABLE III: Suggested Patterns

A1	• Investment in savings from 30%-50% • 40%-70% in mma • Marginal of 10% in cds
A2	• Invest upto 10% in cds
A3	• Investment of 10%-30% in savings
A4	• Suggestion of most of the part of the financial assest into savings(30%-80%)
A5	• Minimal amount in stocks and equity upto 10% • Upto 10% in savings

Output details: The output of FP growth algorithm specifies the most probable investment patterns of the people who have same demographic factors as the user in 11 different financial instruments. 11 financial instruments are Savings, CALL (call accounts), Bonds, NMMF (directly held pooled investment funds), MMA (money market accounts), CDS (certificates of deposits), RETEQ (quasi liquid retirement accounts), Stocks, Saving Bonds, CashLI (cash value of whole life insurance), and Equity. All the percentages in the results were stated with financial assets of the user's input set as the base.

VI. Conclusion and Future Work

The method for extraction of investment patterns among the large investment dataset was carried. The results were the series of investment patterns mined out. The results were in the form of categorized percentage investment in the certain financial instruments.

The future work includes the validation of the results from the real time data and analysis. The work on to find the success rate of the extracted results is an area which is being researched. The success rate defined as the results of the similar investments as the suggested one and well versed with the market movements. The success should be the weighted average of the maximum returns and management of the risk associated with the instruments that are suggested for investment. Further progress is being made to make a real time application or a tool to help the investors in finance world in investment, so that they do not depend on the third party people for example brokers for their investment strategies.

Acknowledgement

Sakshi Singh thanks the Computer Science department of Jaypee Institute of Information Technology for the technical support needed for the research. Harsh Mittal thanks his co author Archana Purwar for constantly guiding and giving directions required during the work.

References

[1] J. E. Grable and S. H. Joo, "Environmental and biopsychosocial factors associated with financial risk Tolerance," *Association for Financial Counseling and Planning Educatio*, vol. 15, no. 1, pp. 73-82, 2004.

[2] C. Veld and Y. V. Veld-Merkoulova, "The risk perceptions of individual investors," *Journal of Economic Psychology*, vol. 29, issue 2, pp. 226-252, April 2008.

[3] T. Dohmen, A. Falk, D. Huffman, U. Sunde, J. Schupp, and G. G. Wagner, "Individual risk attitudes: measurement, determinants and behavioral Consequences," *Journal of the European Economic Association*, vol. 5517, issue 1997, pp. 522-550, 2009.

[4] N. Nicholson *et al*., "Risk propensity and personality," London Business School, 2002.

[5] W. F. Sharpe, "Capital asset prices: a theory of market equilibrium under conditions of risk," *Journal of Finance*, vol. 19, no. 3, pp. 425-442, September 1964.

[6] S. A. Ross, "The capital asset pricing model (CAPM), short-sale restrictions and related issues," *Journal of Finance*, vol. 32, no. 177, 1977.

[7] C. Soares, Y. Peng, J. Mengc, T. Washio, and Z. H. Zhou, "Applications of data mining in e-business and finance: introduction," presented at the 2008 Conference on Applications of Data Mining in E-Business and Finance, The Netherlands, 2008.

[8] P. C. H. Ma and K. C. C. Chan, "Incremental fuzzy mining of gene expression data for gene function prediction," *IEEE Transactions on Biomedical Engineering*, vol. 58, issue 5, pp. 1246-1252, May 5, 2011.

[9] L. Tari, C. Baral, and S. Kim, "Fuzzy c-means clustering with prior biological knowledge," *Journal of Biomedical Informatics*, vol. 42, issue 1, pp. 74-81, January 22, 2008.
[10] H. C. Han, D. Xin, and X. Yan, "Frequent pattern mining: current status and future directions," *Data Mining and Knowledge Discovery*, vol. 15, issue 1, pp. 55-86, January 2007.
[11] U. Malmendier and S. Nagel. (2011). Depression babies: do macroeconomic experiences affect risk taking? *The Quarterly Journal of Economics.* [Online]. 126(1). pp. 373-416. Available: http://www.nber.org/papers/w14813
[12] 2007 Survey of consumer finances. [Online]. Available: http://federalreserve.gov/econresdata/scf/scf_2007survey.htm.
[13] H. Jing, "Application of fuzzy data ming algorithm in performance evaluation of human resource," in *Proc. International Forum on Computer Science-Technology and Applications*, December 2009, pp. 343-346.
[14] R. J. Hathaway and J. C. Bezdek, "Fuzzy c-means clustering of incomplete data," *IEEE Transactions on Systems, Man and Cybernetics—Part B: Cybernetics*, vol. 31, issue 5, pp. 735-744, October 2001.

Sakshi Singh was born in Mainpuri, India on 29th September 1989. She is a B.Tech. in Computer Science from Jaypee Institute of Information Technology, Noida, India. She graduated in 2012. Her major field of study was information retrieval and data mining and her interest lies in financial services.

She has worked in LG Electronics as a summer intern and is presently working in SAP Technologies India Pvt. Ltd., Bengaluru, India.

Harsh Mittal was born in Muzaffar Nagar, India on 5th August 1990. He is a B.Tech. in computer science from Jaypee Institute of Information Technology, Noida, India. He graduated in the year of 2012. His major field of study was information retrieval and data mining and his interest lies in application development for mobile technologies.

He has worked in Eazeworks, Noida as a summer intern and is presently working as a systems engineer in Infosys Ltd., Mysuru, India.

Archana Purwar has got her M.Tech.(CSE), MCA, B.Sc(PCM) and has a work experience of 4 years and 5 months. Her interest lies in analysis and design, DBMS, distributed computing, computer organization. She is currently working as a lecturer in Jaypee Institute of Information Technology, Noida, India.

Mrs. Purwar previous publications were: (1) M. Gupta, M. Dave, and A. Purwar, "Simulation of power efficient region based approach for query processing in wireless sensor networks," in *Proc. Second International Conference on Emerging Trends in Engineering & Technology*, pp. 1104-1109, 2009. (2) A. Purwar and R. Nath, "Grid computing architecture and issues related to it," in *Proc. National Conference on Information Technology*, Radaur, Harayana, 2007.