IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Issue-4, April 2019 E-ISSN: 2347-2693

Precision Clustering Based on Boundary Region Analysis for Share Market Database

M. Aruna[1]*, S. Sugumaran[2], V. Srinivasan[3]

[1]Department of Computer Applications, Velalar College of Engineering and Technology, Erode, Tamilnadu, India.
[2]Department of Computer Science, Erode Arts and Science College, Erode, Tamilnadu, India.
[3]Department of Computer Applications, Velalar College of Engineering and Technology, Erode, Tamilnadu, India.

Corresponding Author: arunasrini2005@gmail.com Tel.: +91-9488021443

DOI: https://doi.org/10.26438/ijcse/v7i4.113118 | Available online at: www.ijcseonline.org

Accepted: 15/Apr/2019, Published: 30/Apr/2019

Abstract - In many research areas it's always found that it is very difficult to cluster the databases which come under the close region of clusters. When the database has a unique cluster then it is faster to make the cluster in lesser times but when it is coming to closer region of two or more clusters then the time taken for the clustering is high and need to be clustered very carefully by examining each attributes. In this paper clustering is done using the partitioning method and complex regions are selected which are closed to two or more cluster and this selected database is again carefully examined by each of the attribute and then finally clustered to produce more accuracy than the partitioning method.

Keywords— Boundary region analysis, Precision Clusters, Share market database, Large database, Reduced Dataset, Attribute Selection, etc.

I. INTRODUCTION

The analysis of large database is a central challenge in computer applications especially for complex database. Complex database provides a crucial and well timed contribution for allowing the creation of real time applications that deal with large data of high complexity in which mining on the fly can make an immeasurable difference [1], such as supporting cancer diagnosis, detecting deforestation, share market analysis etc.

II. RELATED WORK

Analysis map number of clusters against number of variables. Then test for efficiency of clustering. The contention is that, given a fixed number of variables, one of them being historic volatility of NIFTY returns, if increase in the number of clusters improves clustering efficiency, then volatility cannot be predicted. Volatility then becomes random as, for a given time period, it gets classified in various clusters [2] [3]. On the other hand, if efficiency falls with increase in the number of clusters, then volatility can be predicted as there is some homogeneity in the data. If fix the number of clusters and then increase the number of variables, this should have some impact on clustering efficiency.

Data mining approach for classification of stocks into clusters After classification, the stocks could be selected from these groups for building a portfolio. It meets the criterion of minimizing the risk by diversification of a portfolio. The clustering approach categorizes stocks on certain investment criteria [4]. Also have used stock returns at different times along with their valuation ratios from the stocks of Bombay Stock Exchange for the fiscal year 2007–2008. Results of our analysis show that K-means cluster analysis builds the most compact clusters as compared to SOM and Fuzzy C-means for stock classification data. And then select stocks from the clusters to build a portfolio, minimizing portfolio risk and compare the returns with that of the benchmark index, i.e. Sensex.

A new methodology for performing a structured cluster analysis of stock market dataset that a tree-based neural network (TTOSOM). The TTOSOM performs self-organization to construct tree-based clusters of vector data in the multi-dimensional space [5]. The resultant tree possesses interesting mathematical properties such as a succinct representation of the original data distribution, and a preservation of the underlying topology. In order to demonstrate the capabilities of TTOSOM method, analyze 206 assets of the Italian stock market. Also were able to establish topological relationships between various companies traded on the Italian stock market and visually inspect the resultant taxonomy. The results are obtained

briefly reported were amazingly accurate and reflected the real-life relationships between the stocks.

III. APPROXIMATIONS

The starting point of approximation is the indiscernibility relation, generated by information concerning objects of interest. The indiscernibility relation is intended to express the fact it is unable to discriminate some objects employing the available information due to the lack of knowledge. An approximation is also other significant notion in Rough Sets Theory, being related with the meaning of the approximations topological operations. The lower and the upper approximations of a set are interior and closure operations in a topology generated by the indiscernibility relation. Below is presented and described the types of approximations.

A. Lower approximation (B")

Description of the domain objects that are known with certainty to belong to the subset of interest represents the lower approximation. The Lower Approximation Set of a set X, with regard to R is the set of all of objects, which certainly can be clustered with X regarding R, that is, set B".

B. Upper Approximation (B)*

Upper Approximation is a description of the objects that probably belong to the subset of interest. The upper approximation set of a set X regarding R is the set of all of objects which can be possibly classified with X regarding R, that is, set B*.

Boundary Region is description of the objects that of a set X regarding R is the set of all the objects, which is difficult to cluster neither as X nor -X regarding R. The boundary region is a placed as $X = \emptyset$ (Empty), then the set is measured as "Crisp", that is, exact in relation to R; otherwise, it is set as $X \neq \emptyset$ (empty) it is considered as X "Rough". The boundary region is defined as BR = B* - B.

In this paper first we find out the boundary region for all the clusters which are difficult to cluster by partioning method and need to examine this region with the help of the attributes selected. The selection of attributes also plays an important role in clustering of the boundary region.

IV. METHODOLOGY

The ability to apply attribute selection is critical for effective analysis, because datasets frequently contain far more information than is needed to build the model. Entropy is one of the simple and fastest attribute selection methods and is often used in clustering techniques [6] [7]. For example, a dataset might contain 100 columns that describe the characteristics of share, but if the data in some of the columns is very sparse that would gain very slight advantage from adding them to the model. If the analysis keeps the unneeded columns while building the model, more CPU and memory are required during the training process, and more storage space is required for the completed model. One should typically need to remove unneeded columns because they might degrade the quality of discovered patterns [8].

This paper uses shennon's entropy for selection of attributes, the share market dataset has more than 25 attributes from which few attribute is select for faster clustering and to produce accuracy result.

A. Shannon's Entropy

In this paper Shannon's Entropy is used which is a fastest attribute ranking methods and is often used in clustering techniques [9]. The proposed system need to select the attribute which gives more information for clustering from a large number of attributes, to obtain the accurate clustering. The gain Entropy is obtained from Information Theory [10]. Equations 1 and 2 are used to calculate the entropy and the information gain. As the entropy and the information gain is calculated, we select the threshold point for the selection of the attributes if the dataset contains more than five attributes and this are order from the highest values [11]. The attribute that contains the highest values are considered for the clustering process by leaving other attributes which does not give more information for the evaluation process [12]. In addition to entropy and the information gain a new method is used to select the attributes, if the attribute contains same value for the ranking, only one attribute is selected for the attributes which have the same value. The section below shows how the attribute selection for share market database is done using entropy and the information gain for the evaluation process.

B. Entropy

Entropy is a measure of variability in a random variable. It will measure how the particular attribute divides the training examples into the number of result classes [13] [14]. Table 1 show the sample dataset for the share market data. In our problem one need to select the attribute which gives more information for clustering so as to make accurate clustering. For defining gain Entropy is obtained from Information Theory. Entropy is used to calculate the amount of useful information in an attribute. This is calculated as Equation. (1)

$$\text{Entropy}(S) = -\Sigma\, P(x_i) \log_b P(x_i) \quad (1)$$

Where:

S = Collection of Samples

x_i = Set of outcomes

$P(x_i)$ = Proportion of S to the class x_i

C. Information Gain (IG)

The information gain relies on the decrease in entropy once a dataset is split on associate attribute. First the attribute that creates the most homogeneous branches are identified Equation (2)

$$IG(Y/X) = H(Y) – H(Y/X) \quad (2)$$

The entropy calculated for the dataset of share market dataset are ordered as public share holding with 0.97, high-average 0.92, 52 weeks high 0.88, Average of 52 weeks 0.86 and so on.

Table 1. Sample Database of Partial Record is shown from the Indian Share Market Dataset

S.No	A1	A2	A3	…	A4	A15
1	Apple Finance Ltd	1.00	1.00		1.00	0.96
2	Brilliant Securities Ltd	0.04	0.05		0.04	0.58
	…		…			
49	Suryanagri Fin Lease Ltd.	0.04	0.02		0.04	0.83
50	Suryavanshi Spinning Mills Ltd.	0.05	0.02		0.05	0.47

A1-S.No, A2-Share name, A3-Q1-Profit from April to June, A-Yearly profit, A5-52_weaks high, A6-Average of q1 to q4, A7-Average profit of the year, A8-Average of 52 weeks, A9-Q2- profit from July to September, A10-Q3-Profit from October to December, A11-Q4-Profit from January to March, A12-Public share holding, A13-52 weeks low, A14-High-Average, A15-Average total, A16-ClassA14-High-Average, A15-Average total.

The selected attributes alone is taken for further process in clustering the share market dataset into three categories of high, Medium and low.

D. Clustering Methods

The process of grouping a set of data sets into multiple teams or groups in order that objects within a group have high similarity, but are very dissimilar to objects in other groups is referred as clustering. Dissimilarities and similarities are assessed based on the attribute value, users describing the objects and often involve distance measures [15]. Clustering as a data mining tool has its roots in many application areas such as biology, security, business intelligence, share market and web search etc.

The search for clusters represents unsupervised learning and from a machine learning perspective clusters correspond to hidden patterns and the resulting system represents a data concept [16]. In data mining applications such as scientific data exploration, information retrieval and text mining, spatial database applications, Web analysis, CRM, marketing, medical diagnostics, computational biology, share market and many others clustering plays an extraordinary role [6]. Clustering is the theme of dynamic research in several fields such as statistics, pattern recognition, and machine learning. Data mining adds to cluster the complications of very large datasets with various attributes of dissimilar types [17]. This enforces distinctive computational requirements on relevant clustering algorithms. There are different clustering methods evolved they are partitioning methods, hierarchical methods, density based methods and grid based methods.

E. Partitioning methods

Partitioning is the simplest and most essential version of cluster analysis in which it organizes the objects of a set into several exclusive groups or clusters. To keep the problem specification concise, one can assume that the number of clusters is given as background knowledge which is considered as starting parameter partitioning methods. Formally, given a data set, D, of n objects, and k clusters, the number of clusters to form, a partitioning algorithm organizes the objects into k partitions (k<=n), where each partition represents a cluster [18]. The clusters are formed to optimize an objective portioning criterion, such as a dissimilarity function based on distance, so that the objects within a cluster are similar to one another and dissimilar to objects in other clusters in terms of the data set attributes. The commonly used partitioning methods are K-means and K-medoids.

F. Hierarchical Methods

A hierarchical clustering method works by grouping data objects into a hierarchy or tree of clusters. Representing data objects in the form of a hierarchy is useful for data summarization and visualization. This builds a cluster hierarchy or, a tree of clusters, also known as a dendrogram. Every cluster node contains child clusters; sibling clusters partition the points covered by their common parent. Such an approach allows exploring 6 data on different levels of granularity [19]. Hierarchical clustering methods are categorized into agglomerative (bottom-up) and Divisive (top-down). An agglomerative clustering starts with one point (singleton) clusters and recursively merges two or more most appropriate clusters. A divisive clustering starts with one cluster of all data points and recursively splits the

most appropriate cluster. The process continues until a stopping criterion is achieved.

G. Grid-Based Methods

A grid based clustering method takes a space driven approach by partitioning the embedding space into cells independent of the distribution of the input objects. The grid based clustering approach uses a multi resolution grid data structure. It quantizes the object space into a finite number of cells that form a grid structure on which all of the operations for clustering are performed. The main advantage of the approach is its fast processing time, which is typically independent of the number of data object. The methods used for density based clustering are STING and CLIQUE.

H. Density-Based Methods

To find the clusters of arbitrary shape such as the "S" shaper and oval clusters. They would likely inaccurately identify convex regions, where noise or outliers are included in the clusters. An open set in the Euclidean space can be divided into a set of its connected components. The implementation of this idea for partitioning of a finite set of points requires concepts of density, connectivity and boundary. They are closely related to a point is nearest neighbors. A cluster, defined as a connected dense component, grows in any direction that density leads. Therefore, density-based algorithms are capable of discovering clusters of arbitrary shapes. Also this provides a natural protection against outliers. The basic techniques of density based clustering methods are DBSCAN, OPTICS and DENCLUE.

In this paper the partitioning method is used for the share market database to cluster as high, medium and low.

V. PROPOSED MODEL

Share market dataset was framed from getting the information from the National Stock Exchange (NSE) is India's leading stock exchange covering various cities and towns across the country. NSE was set up by leading institutions to provide a modern, fully automated screen-based trading system with national reach [20]. The Exchange has brought about unparalleled transparency, speed & efficiency, safety and market integrity. It has set up facilities that serve as a model for the securities industry in terms of systems, practices and procedures.

This database was taken to show whether a particular share will raise or go down in future coming year based on the previous year recorded information [21]. The different attributes for the dataset are Q1 profit of April to June, Q2 profit of July to September, Q3 profit of October to December, Q4 profit of January to March, Yearly profit, Public share holding, 52 Weeks Low, 52 Weeks High, Average Profit of Q1 to Q4, Average profit of the Year, Average of 52 weeks and so on. To reduce the attribute, Entropy and the information gain are calculated for each attribute with the equation (1) and (2). These attributes are then ordered according to the highest entropy and the information gain [22] [23]. Finally selects only few attributes for the final clustering process of complex dataset. In this paper the share market dataset which has 16 attribute is reduced to four attributes which gives meaningful information for clustering to be classified as high, medium and low leaving other attributes that does not give more information for the clustering purpose. The selected attributes are listed in the table below this is done based on the entropy and information gain [24].

Table 2. Selected Share Market Dataset

S. No	A6	A2	A7	A8	Total
1	11.278	0.065	0.055	0.047	11.446
2	23.545	0.304	0.265	0.265	24.379
3	7.947	12.606	12.298	12.330	45.181
	…		…		…
48	14.719	0.815	0.976	0.978	17.487
49	20.674	1.460	1.048	0.985	24.167
50	11.232	0.144	0.358	0.240	11.974

A6 – Public share holding
A2 – Yearly profit
A7 – Average profit of the year
A8 – 52_weaks high

The above share market database is clustered as high, medium and low based on the weight and selecting the center point based on the mean of the weight. Below diagram shows the three different clustering of share market dataset to predict low, medium and high [25] [26].

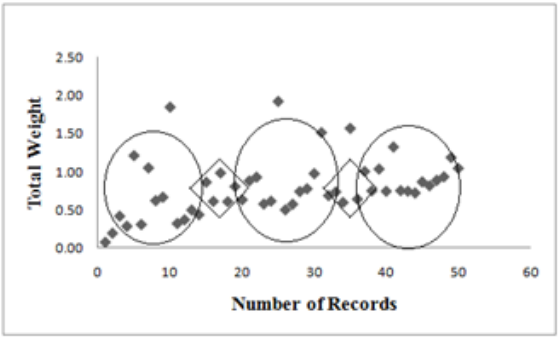

Figure 1. Boundary Region Analysis

The diagram shows that the clustering is done for the given share market database as low, medium and high. In this

paper we argue that the boundary regions near the clusters may not be accurately clustered and one need to analysis the boundary region before the cluster is made. This can be done by analyzing the each attribute and then cluster, as mean weight of the attribute may not cluster the share correctly in the boundary region. So in this approach each attribute field is examined before the cluster is made, this boundary region analysis give the precision result for the cluster.

The share market that falls in the boundary region is alone taken and these shares are clustered based on examining each attribute by using the if-else statement. Taking each of the shares that falls in the boundary region and analyzed carefully for first three attribute to be clustered into accurate groups.

Table 3. Ordered Share Market Dataset

S. No	A6	A2	A7
1	0.35	0.19	0.16
2	0.36	0.13	0.07
3	0.37	0.28	0.17
4	0.39	0.11	0.06
5	0.39	0.24	0.09
6	0.46	0.15	0.13
7	0.52	0.07	0.05
8	0.53	0.07	0.07
9	0.54	0.03	0.01
10	0.54	0.05	0.02

The selected share is carefully examined by the rules given below and based on the rules the shares are clustered accurately. Out of 5 shares from the cluster low and medium selected as boundary region three are clustered as low and other 2 is clustered as medium cluster. From the 5 shares of medium and high 4 is grouped into medium and 1 is grouped in the high cluster. This is done by the average of center point for the boundary region.

Rule 1: If A1 is L and A2 is L and A3 is L then R is L
Rule 2: If A1 is L and A2 is L and A3 is H then R is L
Rule 3: If A1 is L and A2 is H and A3 is L then R is L
Rule 4: If A1 is L and A2 is H and A3 is H than R is H
Rule 5: If A1 is H and A2 is L and A3 is L then R is L
Rule 6: If A1 is H and A2 is L and A3 is H then R is H
Rule 7: If A1 is H and A2 is H and A3 is L then R is H
Rule 8: If A1 is H and A2 is H and A3 is H then R is H
L – Low, H – High

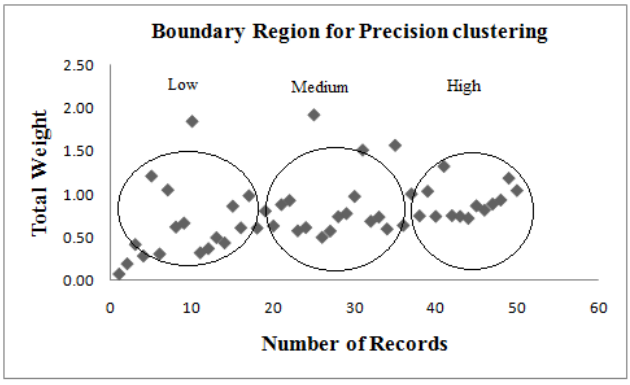

Figure 2. Boundary Region for Precision Clustering

The above diagram shows the clusters made after the selection of the boundary region and applied the rules for the selected shares of the boundary region and finally clustered accurately as low medium and high.

VI. RESULT AND CONCLUSION

The share market database that was clustered based on the partition method may not have the precision cluster as there might be some shares that may not be correctly clustered due to the nearness of the different cluster region. The selected attribute in the boundary region is again clustered by examining each of the attribute for their information and this is done based on the if-then rules that is if any of the two attribute has the same value as high then it is clustered to the high cluster if any of the two attribute has the same value as low then it is clustered as low clusters this is repeated for the other two cluster comparison and clustered based on the if-then rules. So by analyzing the share that fall in the boundary region by checking with the each attribute of the information and then cluster the shares which give the precision result than the partitioning methods.

REFERENCES

[1] Dingsheng, W, Xiang, R, & Yuting, H 2010, 'Data mining algorithmic research and application based on information entropy', Pattern Recognition, vol. 43, no. 1, pp. 5–13.

[2] Gheyas, I & Smith, L 2010, 'Feature subset selection in large dimensionality domains', Pattern Recognition, vol. 43, no. 1, pp. 5-13.

[3] V.Srinivasan 'Feature Selection Algorithm using Fuzzy Rough Sets for Predicting Cervical Cancer Risks' International Journal Modern Applied Science Vol. 4 Issue 9 10 sep 2010.

[4] Guyon, I & Elisseeff, A 2003, 'An introduction to variable and feature selection', Journal of Machine Learning Research, vol. 3, no. 7, pp. 1157-1182.

[5] Halperin, E & Karp, RM 2005, 'The minimum-entropy set cover problem', Theoretical Computer Science, vol. 348, no. 2, pp. 240-250.

[6] Indian Stock Exchange 2003, Stock Board information. Available from <http:// www.stocksabroad.com> [20 June 2003].

[7] Indian Stock Exchange 2005, Share Market Information. Available from <http://www.nse-india.com> [10 May 2005].

[8] V.Srinivasan 'Classify the student with missing value to calculate future semester result for placement record using knowledge acquisition' National Journal Vol. 3 Issue No.3 2010.

[9] Jiawei, H & Micheline, K 2006, 'Data Mining: Concepts and Techniques, 2nd edition', Morgan Kaufmann Publishers.

[10] Kedarnath, JB. & Nur, AT 2007, 'Relationship between entropy and test data compression', IEEE Transaction on Computer Aided Design of Integrated Circuits and Systems, vol. 26, no. 2, pp. 386-395.

[11] V.Srinivasan 'A fuzzy fast classification for share market database with lower and upper bounds' American journal of Applied Science, vol.12, sep 2012, PP.1934-1939.

[12] Liu, H 2005, 'Evolving feature selection', IEEE Intelligence System, vol. 20, no. 6, pp. 64-76.

[13] Neelima, B, Jha, CK & Sandeep KB 2012, 'Application of Neural Network in Analysis of Stock Market Prediction', International Journal of Computer Science & Engineering Technology, vol. 3, no. 4, pp. 61-68.

[14] Richard, J & Qiang, S 2007, 'Fuzzy-Rough Sets Assisted Attribute Selection', IEEE Transactions on Fuzzy Systems, vol. 15, no. 1, pp. 73-89.

[15] V.Srinivasan 'Fuzzy Classification to Classify the Income Category Based on Entropy' Journal of Computer science and Technology Vol 11 No.2, 2011.

[16] Slezak, D 2002, 'Approximate Entropy Reducts', Fundamenta Informaticae, vol. 53, no. 3, pp. 365-390.

[17] Turiel, A & Vicente, CJP 2003, 'Multifractal geometry in stock market time series', Physica A: Statistical Mechanics and its Applications, vol. 322, no. 1, pp. 629-649.

[18] V.Srinivasan 'Fuzzy Fast Classification Algorithm with Hybrid of ID3 and SVM' International journal of Intelligence and Fuzzy System, Vol.24, May 2013, pp.556-561.

[19] Yucel, S, Arnold R & YunTong, W 2004, 'Value of Information Gained From Data Mining in the Context of Information Sharing', IEEE Engineering Management, vol. 51, no. 4, pp. 441-450.

[20] Zabir, HK, Tasnim, SA & Md, AH 2011, 'Price Prediction of Share Market using Artificial Neural Network', International Journal of Computer Applications, vol. 22, no. 2, pp. 0975–8887.

[21] V.Srinivasan 'A Fuzzy Approach to Replace the Missing Data in Large Dataset', International Journal of Applied Engineering and Research, Vol.10, N0.38, May 2015, pp.28312-28317.

[22] M. Setnes, 'Supervised Fuzzy Clustering for Rule Extraction,' IEEE Trans. Fuzzy Systems, vol. 8, pp. 416-424, 2000.

[23] Sathyamoorthy. S, 'Data Mining and Information Security in Big Data,' International Journal of Scientific Research in Computer Science and Engineering, vol. 5, No.3, pp. 86-91, June 2017.

[24] Mantripatjit Kaur, Anjum Mohd Aslam, 'Big Data Analytics on IOT: Challenges, Open Research Issues and Tools", International Journal of Scientific Research in Computer Science and Engineering, vol. 6, No.3, pp. 81-85, June 2018.

[25] Manju Bhardwaj, 'Faculty Link Detection in Cluster based Energy Efficient Wireless Sensor Networks", International Journal of Scientific Research in Network Security and Communications, vol. 5, No.3, pp. 81-85, June 2017.

[26] Bhupendra Kumar Jain, Manish Tiwari, 'Prediction Analysis Technique based on Clustering and Classification', International Journal of Computer Sciences and Engineering, Vol.6, No.6, June 2018.

Authors Profile

M.Aruna completed her MCA in 2005 from Anna university and M.Phil in the year 2008 from Annamalai University . She is currently pursuing Ph.D. and working as Associate Professor in Department the Department of Computer Applications, Velalar College of Engineering and Technology, Erode since 2006. She is a life member of IEEE. She has published around 5 Research papers in reputed international journals and have presented papers in National and International conferences . Her main research work focuses on Data Mining, Clustering, Feature Selection and Big Data Analytics. She has 12 years of teaching experience and 6 years of Research Experience.

S.Sukumaran Associate Professor of computer Science, Erode Arts and Science College has 21 years of research experience and about 30 years of teaching experience. He has published more than 80 papers in National and International Journals. He has presented around 31 papers in National Conferences and 25 Papers in International Conferences. He has received funds from UGC for minor research project. He has guided 13 Ph.D scholars and 6 are in-progress .He is a reviewer for various journals. His area of specialization includes Image Processing and Data Mining.

V. Srinivasan pursed Bachelor of Science from Bharathidasan College of Arts and Sceince, India in 1996 and Master of Computer Applications from Kongu Engineering College, India in year 2002 and Completed his Ph.D in Anna Universtiy , India in year 2014 and currently working as Professor in Department of Computer Applications, Velalar College of Engineering and Technology, India since 2006. He is a life member of IEEE & IEEE computer society since 2009. He has published more than 6 research papers and 20 conferences. His main research work focuses on Data Mining, Fuzzy Classification, Feature Selection, Big Data Analytics. He has 16 years of teaching experience and 4 years of Research Experience.

