International Journal of Enterprise Computing and Business Systems
ISSN (Online) : 2230-8849
Volume 4 Issue 2 July 2014
International Manuscript ID : ISSN22308849-V4I2M51-072014

AN EFFECTIVE DYNAMIC UNSUPERVISED CLUSTERING ALGORITHMIC APPROACH FOR MARKET BASKET ANALYSIS

Sheenu Verma
M.Tech. Research Scholar
Ambala College of Engineering and Applied Research
Mithapur, Haryana, India

Sakshi Bhatnagar
Assistant Professor
Ambala College of Engineering and Applied Research
Mithapur, Haryana, India

ABSTRACT

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters). It is a main task of exploratory data mining, and a common technique for statistical data analysis, used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, and bioinformatics. Cluster analysis itself is not one specific algorithm, but the general task to be solved. It can be achieved by various algorithms that differ significantly in their notion of what constitutes a cluster and how to efficiently find them. Popular notions of clusters include groups with small distances among the cluster members, dense areas of the data space, intervals or particular statistical distributions. Clustering can therefore be formulated as a multi-objective optimization problem. The appropriate clustering algorithm and parameter settings (including values such as the distance function to use, a density threshold or the number of expected clusters) depend on the individual data set and intended use of the results. Cluster analysis as such is not an automatic task, but an iterative process of knowledge discovery or interactive multi-objective optimization that involves trial and failure. It will often be necessary to modify data preprocessing and model parameters until the result achieves the desired properties. Besides the term clustering, there are a number of terms with similar meanings, including automatic classification, numerical taxonomy, botryology (from Greek βότρυς "grape") and typological analysis. The subtle differences are often in the usage of the results: while in data mining, the resulting groups are the matter of interest, in automatic classification the resulting discriminative power is of interest. This often leads to misunderstandings between

International Journal of Enterprise Computing and Business Systems
ISSN (Online) : 2230-8849
Volume 4 Issue 2 July 2014
International Manuscript ID : ISSN22308849-V4I2M51-072014

researchers coming from the fields of data mining and machine learning, since they use the same terms and often the same algorithms, but have different goals. In this research work, the new clustering algorithm is proposed and implemented for the fuzzy based implementation of dynamic clusters of market basket analysis.

Keywords – Data Mining, Clustering, Knowledge Discovery in Databases

Data Mining

Data mining [9] refers to the analysis of the large quantities of data that are stored in computers. Data mining has been called exploratory data analysis, among other things. Masses of data generated from cash registers, from scanning, from topic specific databases throughout the company, are explored, analyzed, reduced, and reused. Searches are performed across different models proposed for predicting sales, marketing response, and profit. Classical statistical approaches are fundamental to data mining. Automated AI methods are also used. Data mining requires identification of a problem, along with collection of data that can lead to better understanding and computer models to provide statistical or other means of analysis [8].

Data comes in, possibly from many sources. It is integrated and placed in some common data store. Part of it is then taken and pre-processed into a standard format. This 'prepared data' is then passed to a data mining algorithm which produces an output in the form of rules or some other kind of 'patterns'[18].

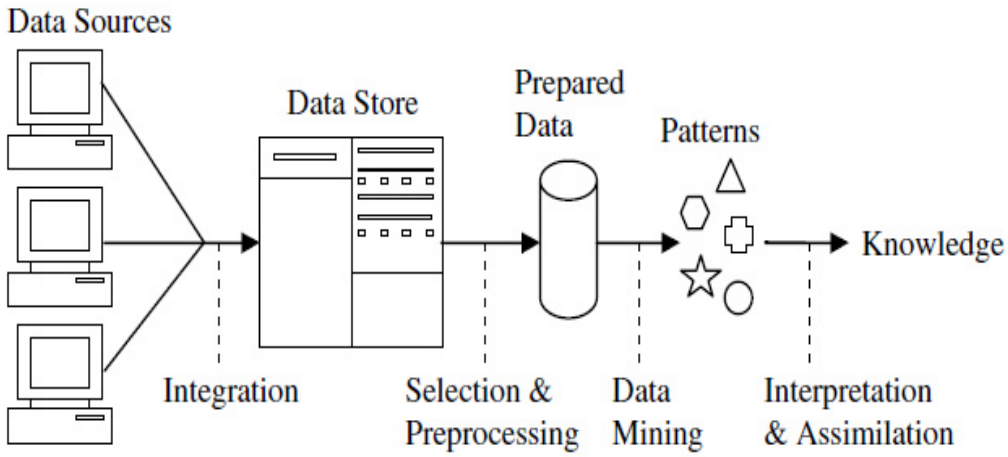

Figure 1.1: The Knowledge Discovery Process [18]

A variety of analytic computer models have been used in data mining. The standard model types in data mining include regression (normal regression for prediction, logistic regression for classification), neural networks, and decision trees. These techniques are well known [8].

Data Mining Requisites

Information mining requires distinguishing proof of an issue, on top of gathering of information that can accelerate better comprehension, and machine models to give measurable or different method of investigation. This may be backed by visualization instruments, that showcase information, or through basic measurable investigation, for example association examination. Information mining devices need to be adaptable, adaptable, fit for exactly expecting reactions between movements and results, and equipped for programmed usage. Flexible implies the capacity of the instrument to apply a wide mixed bag of models. Versatile devices intimate that if the devices deals with a little information set, it may as well likewise finish up bigger information sets. Mechanization is handy, however its requisition is relative. Some diagnostic capacities are frequently mechanized, however human setup preceding bringing about techniques is needed. Truth be told, investigator judgment is basic to efficacious usage of information mining. Fitting determination of information to incorporate in inquiries is discriminating. Information conversion additionally is frequently needed. An excessive amount of variables transform a lot of yield, while excessively few can neglect enter relationships in the information. Key comprehension of factual notions is required for fruitful information mining.

Clustering

Clustering is an important KDD technique with numerous applications, such as marketing and customer segmentation. Clustering typically groups data into sets in such a way that the intra-cluster similarity is maximized and while inter-cluster similarity is minimized [11]. Clustering is an unsupervised learning. Clustering algorithms examines data to find groups of items that are similar. For example, an insurance company might group customers according to income, age, types of policy purchased, prior claims experience in a fault diagnosis application, electrical faults might be grouped according to the values of certain key variables [18].

International Journal of Enterprise Computing and Business Systems
ISSN (Online) : 2230-8849
Volume 4 Issue 2 July 2014
International Manuscript ID : ISSN22308849-V4I2M51-072014

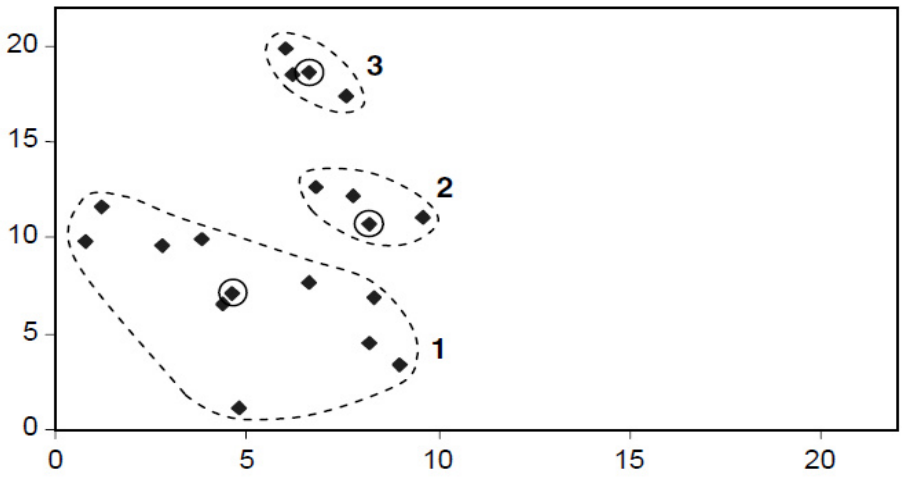

Figure 1.2: Clustering of Data [18]

Most previous clustering algorithms focus on numerical data whose inherent geometric properties can be exploited naturally to define distance functions between data points. However, much of the data existed in the databases is categorical, where attribute values can't be naturally ordered as numerical values. Due to the special properties of categorical attributes, the clustering of categorical data seems more complicated than that of numerical data [13]. To overcome this problem, several data-driven similarity measures have been proposed for categorical data. The behaviour of such measures directly depends on the data [24].

Grouping might be recognized the most essential unsupervised studying issue; thus, as each other issue of this kind, it manages discovering a structure in an accumulation of unlabeled information. A detached meaning of grouping could be "the procedure of forming questions into gatherings whose parts are comparative somehow". A bunch is thusly an accumulation of articles which are "comparative" between them and are "different" to the items fitting in with different bunches. We can demonstrate this with a straightforward graphical case:

International Journal of Enterprise Computing and Business Systems
ISSN (Online) : 2230-8849
Volume 4 Issue 2 July 2014
International Manuscript ID : ISSN22308849-V4I2M51-072014

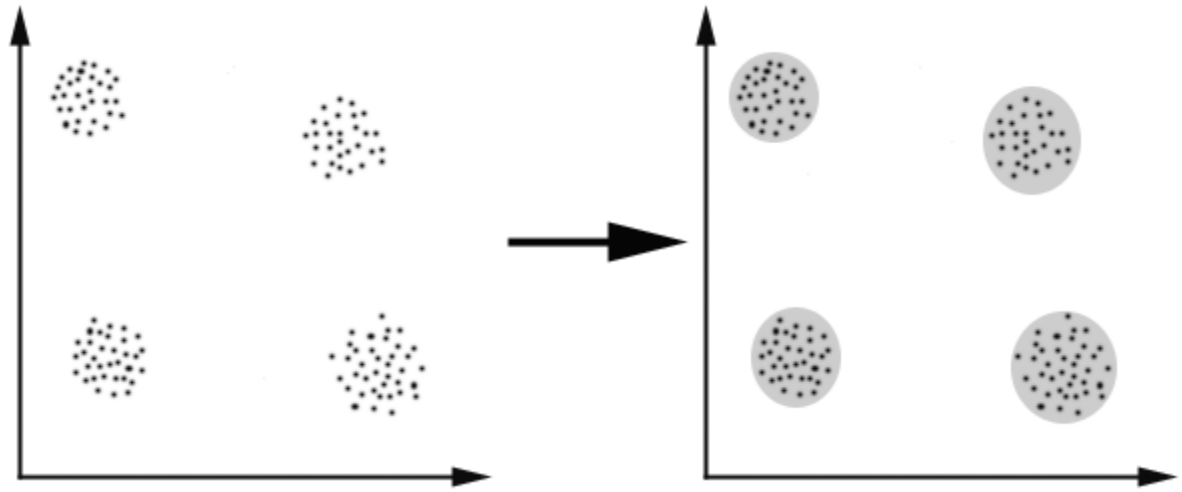

Figure 1.3: Clustering

In this case we easily identify the 4 clusters into which the data can be divided; the similarity criterion is distance: two or more objects belong to the same cluster if they are “close” according to a given distance (in this case geometrical distance).This is called distance-based clustering.

A different sort of grouping is reasonable clustering: two or more questions have a place with the same bunch if this one describes an idea regular to all that protests. In different expressions, articles are gathered as per their fit to enlightening ideas, not as per modest likeness measures.

Goal of Clustering

The objective of clustering is to figure out the innate amassing in a set of unlabeled information. Yet how to choose what constitutes a great grouping? It could be indicated that there is no supreme "best" rule which might be free of the last point of the bunching. Thus, it is the client which should supply this basis, in such a path, to the point that the aftereffect of the grouping will suit their necessities. Case in point, we could be fascinated by finding delegates for homogeneous gathers (information decrease), in finding "characteristic groups" and portray their obscure lands ("common" information sorts), in finding of service and suitable groupings ("helpful" information classes) or in finding irregular information objects (outlier location).

Classification of Methods

Partitional clustering

Partition-based methods construct the clusters by creating various partitions of the dataset. So, partition gives for each data object the cluster index pi. The user provides the desired number of clusters M, and some criterion function is used in order to evaluate the proposed partition or the solution. This measure of quality could be the average distance between clusters; for instance, some well-known algorithms under this category are k-means, PAM and CLARA. One of the most popular and widely studied clustering methods for objects in Euclidean space is called k-means clustering. Given a set of N data objects xi and an integer M number of clusters. The problem is to determine C, which is a set of M cluster representatives cj, as to minimize the mean squared Euclidean distance from each data object to its nearest centroid [21].

Hierarchical clustering

Hierarchical clustering methods build a cluster hierarchy, i.e. a tree of clusters also known as dendogram. A dendrogram is a tree diagram often used to represent the results of a cluster analysis. Hierarchical clustering methods are categorized into agglomerative (bottom-up) and divisive (top-down) as shown in Figure 4. An agglomerative clustering starts with one-point clusters and recursively merges two or more most appropriate clusters. In contrast, a divisive clustering starts with one cluster of all data points and recursively splits into nonoverlapping clusters [21].

Density-based and grid-based clustering

The key idea of density-based methods is that for each object of a cluster the neighborhood of a given radius has to contain a certain number of objects; i. e. the density in the neighborhood has to exceed some threshold. The shape of a neighborhood is determined by the choice of a distance function for two objects. These algorithms can efficiently separate noise. DBSCAN and DBCLASD are the well-known methods in the density based category. The basic concept of grid-based clustering algorithms is that they quantize the space into a finite number of cells that form a grid structure. And then these algorithms do all the operations on the quantized space. The main advantage of the approach is its fast processing time, which is typically independent of the number of objects, and depends only on the number of grid cells for each dimension. Famous methods in this clustering category are STING and CLIQUE [21].

Outliers

An outlying observation, or outlier [6], is one that appears to deviate markedly from other members of the sample in which it occurs. Outlier points can therefore indicate faulty data, erroneous procedures, or areas where a certain theory might not be valid. However, in large samples, a small number of outliers is to be expected.

International Journal of Enterprise Computing and Business Systems
ISSN (Online) : 2230-8849
Volume 4 Issue 2 July 2014
International Manuscript ID : ISSN22308849-V4I2M51-072014

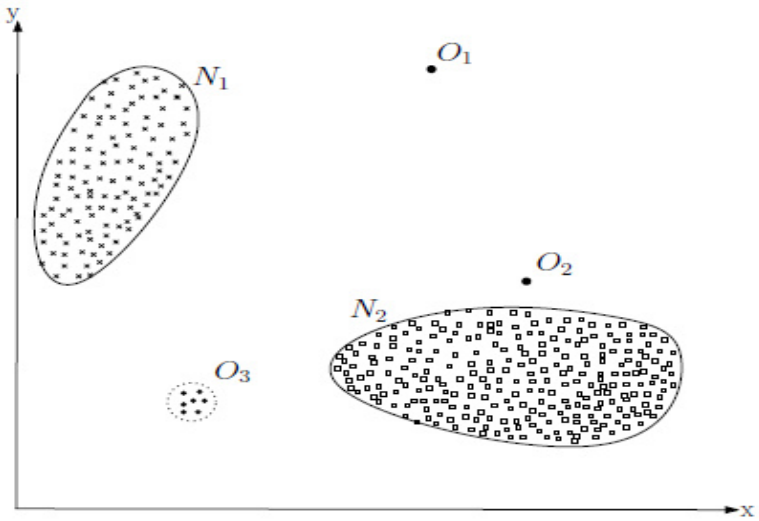

Figure1.4: Outliers in two-dimensional dataset [26]

Figure 1.4 illustrates outliers in a two dimensional dataset. The data has two normal regions, N1 and N2. O1 and O2 are two outlying instances while O3 is an outlying region. The outlier instances are the ones which do not lie within the normal regions [26].

Outlier Detection

Most outlier detection techniques treat objects with K attributes as points in ℜK space and these techniques can be divided into three main categories. The first approach is distance based methods, which distinguish potential outliers from others based on the number of objects in the neighborhood. Distribution-based approach deals with statistical methods that are based on the probabilistic data model. A probabilistic model can be either a priori given or automatically constructed using given data. If the object does not suit the probabilistic model, it is considered to be an outlier. Third, density-based approach detects local outliers based on the local density of an object's neighborhood. These methods use different density estimation strategy. A low local density on the observation is an indication of a possible outlier.

Distance-based approach

In Distance-based methods outlier is defined as an object that is at least dmin distance away from k percentage of objects in the dataset. The problem is then finding appropriate dmin and k such that outliers would be correctly detected with a small number of false detections. This process usually needs domain knowledge.

Definition: A point x in a dataset is an outlier with respect to the parameters k and d, if no more than k points in the dataset are at a distance d or less from x.

To explain the definition by example we take parameter k = 3 and distance d as shown in Figure 1.5. Here are points xi and xj be defined as outliers, because of inside the circle for each point lie no more

than 3 other points. And x′ is an inlier, because it has exceeded number of points inside the circle for given parameters k and d [21].

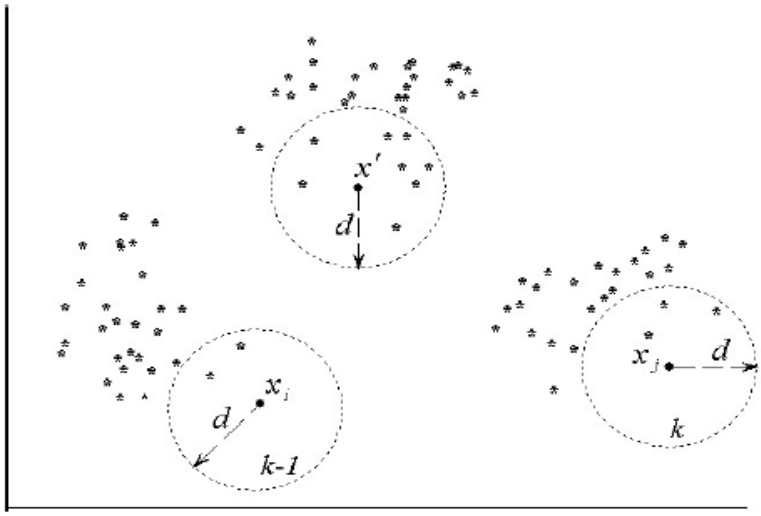

Figure 1.5: Illustration of outlier definition by Knorr and Ng [21]

Distribution-based approach

Distribution-based methods originate from statistics, where object is considered as an outlier if it deviates too much from underlying distribution. For example, in normal distribution outlier is an object whose distance from the average object is three times of the variance [21].

Density-based approach

Density-based methods have been developed for finding outliers in a spatial data. These methods can be grouped into two categories called multi-dimensional metric space-based methods and graph-based methods. In the first category, the definition of spatial neighbourhood is based on Euclidean distance, while in graph-based spatial outlier detections the definition is based on graph connectivity. Whereas distribution-based methods consider just the statistical distribution of attribute values, ignoring the spatial relationships among items, density-based approach consider both attribute values and spatial relationship [21].

Problem Statement

THE PRAGMATIC INVESTIGATION AND COMPARATIVE ANALYSIS IS THE THRUST AREA FOR EFFICIENT AND EFFECTIVE RESULTS OF THE IMPROVED ALGORITHM WITH RESPECT TO THE EXISTING APPROACH

International Journal of Enterprise Computing and Business Systems
ISSN (Online) : 2230-8849
Volume 4 Issue 2 July 2014
International Manuscript ID : ISSN22308849-V4I2M51-072014

If INPUT: $\sum (D_i.\{R(T_i)\}\text{->FF}) \Rightarrow C_R$ ʊ $R(PCT(R):PTG(R)) \Rightarrow CF_m$ m <=i

where D_i= Data repository

$R(T_i)$=Record / Data Set of the Transaction

FF= Fitness Function

C_R=Cluster Eligibility of the Record R

CF_m=Final Eligibility for the Cluster m where m <=i

PTG = Percentage Equivalent of R

PCT = Percentile Equivalent Value of R

then we want to achieve following

OUTPUT: $(\text{Ɛ } C_i) \propto (T_i - T_{i-1})$

where C_i = Cluster set

At the first instance, there is a Data Repository (D). In the Data Repository, there will be number of transactions or data sets or transactional data. Each transaction or record or data set is termed as $R(T_i)$. The Data Set regardless of the value and associated parameters will be eligible and move forward the fitness function modeling. Once the fitness function is applied based on the percentile based measurement, it will form the new criteria for the inclusion in the clusters. Finally the set of clusters will be generated with efficient results and optimal time parameter.

Objectives of our research :

1. To devise and implement a novel and efficient technique for dynamic as well as effective cluster formation.
2. To apply and fetch the meaningful records in form of the aggregate values or clusters for intelligence and predictions.
3. To analyze the proposed cluster formation algorithm with the existing technique and to prove the effectiveness of the proposed work.
4. To devise a novel fitness function to the transactional data so that the eligibility or relevance of the record can be analyzed.

Problem Significance

Clustering is a well-studied data mining problem that has found applications in many areas. For example, clustering can be applied to a document collection to reveal which documents are about the same topic. The objective in any clustering application is to minimize the inter-cluster similarities and maximize the intra-cluster similarities. There are different clustering algorithms each of which may or may not be suited to a particular application. The traditional clustering paradigm pertains to a single dataset. Recently, attention has been drawn to the problem of clustering multiple heterogeneous datasets where the datasets are related but may contain information

about different types of objects and the attributes of the objects in the datasets may differ significantly. A clustering based on related but different object sets may reveal significant information that cannot be obtained by clustering a single dataset.

Existing Methodology

Existing Algorithm repeatedly reads tuples one by one from the dataset. When the first tuple arrives, a new cluster is formed. The consequent tuples are either put in the existing clusters or rejected by the existing clusters to form a new cluster based on the similarity measure between a tuple and a cluster. In existing approach each tuple belongs to one cluster only.

Proposed Methodology

In proposed algorithm, suppose there are n tuples as shown in figure 4.2. A fitness value is assigned to each tuple using the fitness function. Based upon this fitness value the tuples will be assigned to the clusters. If the fitness value of the tuple is equal to or nearly equal to the threshold value of the generated set of random clusters then only the tuple will be assigned to the cluster otherwise tuple is assigned to the outlier cluster. If there are many clusters in the outlier cluster then a similarity is calculated among theses clusters and outlier is detected. In this approach, tie can also occur i.e. if a tuple belongs to two clusters then we can arbitrarily assign this tuple to any one cluster

Pseudocode for Clustering and Outlier Detection

1. CLUSTERING

1. Generate Dataset (Sequentially or Randomly) / Tuple series (from huge data warehouse)
   $|T|= \{ T_i \mid i \epsilon (1,N) \}$
2. Assign Fitness Value (F_i) to each tuple based on the Acceptance / Rejection of the Data Item for joining the Cluster
   $T= \{ T_i[F_i] \mid i \epsilon (1,N) \}$
3. Generate the set of random clusters (if already exists)
   $C= \{ C_i \mid i \epsilon (1,N) \}$ and assign Threshold
4. Compare T with C based on Fitness value and Associated Parameters
5. If (C_i == NULL) AND $T_{fitness}$==NULL GoTo Step 7
   Else
   If (C_i= First Cluster)
   Assign initial threshold based on the application
   Else
   GoTo Step 1
6. End

International Journal of Enterprise Computing and Business Systems
ISSN (Online) : 2230-8849
Volume 4 Issue 2 July 2014
International Manuscript ID : ISSN22308849-V4I2M51-072014

2. OUTLIER DETECTION

1. Read First / Next Data Item from C_i

   If F_i not matching to predefined threshold to maximum extent

   Similarity (F_i | $F(C_i)$)

   Then

   Separate the tuple and mark it as Outlier

2. If Similar Outliers Occurs

   Then

   Create Clusters of these outlier entries

3. Get Statistics
4. End

3. FINAL INVESTIGATION

Generate Final Results from both algorithms and calculate complexity.

Proposed Algorithm for Fitness Function

In the algorithm given below, for calculation of fitness value our first step is to count the occurences of each product id. Then upper bound is calculated based on the percentage of the occurence of each product. Here, TopPercentage is set as upper bound. Further, percentile is calculated of each product based on the upper bound. In the next step of algorithm, clusters are formed with respect to percentile. Thereafter, if difference of threshold and percentile is same, select cluster arbitrarily otherwise tuple goes to the cluster for which difference is minimum. Then, get statistics and performance report.

Pseudocode

1. Count occurrences of each product id/parameters/interest factor
2. Calculation of Upper Bound based on the Percentage of the occurrence

TopPercentage=UB

3. Calculate Percentile of each product/parameter based on the UB
4. Cluster Formation w.r.t percentile
5. If difference is same, select cluster arbitrarily
6. Get statistics and performance report

CLUSTER FORMATION PROCESS BASED ON THE FITNESS FUNCTION VALUES AND THRESHOLD
(Proposed approach)

Cluster 1 **Cluster 2** **Cluster 3**

International Journal of Enterprise Computing and Business Systems
ISSN (Online) : 2230-8849
Volume 4 Issue 2 July 2014
International Manuscript ID : ISSN22308849-V4I2M51-072014

		Product ID (020)
		Product ID (010)
		Product ID (009)
	Product ID (017)	Product ID (007)
PID (012)		
	Product ID (001)	Product ID (008)
		Product ID (014)
		Product ID (006)

CLUSTER FORMATION PROCESS BASED ON THE IMPLEMENTATION OF THE EXISTING ALGORIHTM

		Product ID (012)
		Product ID (017)
No Product meet the Fitness Function Criteria		
		Product ID (001)
		Product ID (003)
		Product ID (015)

International Journal of Enterprise Computing and Business Systems
ISSN (Online) : 2230-8849
Volume 4 Issue 2 July 2014
International Manuscript ID : ISSN22308849-V4I2M51-072014

Product ID (018)

Product ID (004)

Product ID (002)

Product ID (016)

Product ID (019)

Product ID (005)

Product ID (011)

Product ID (013)

Product ID (020)

Product ID (010)

Product ID (009)

Product ID (007)

Product ID (008)

Product ID (014)

Product ID (006)

Execution Time of Proposed Approach= 3.064 microseconds
Execution Time of Existing Approach = 5.071 microseconds

Percentile based implementation (proposed) takes less execution time than existing based implementation.

It is evident from the simulated environment that the classical technique of the cluster formation is generating the clusters in very vague and very traditional method without intelligence and bound based measurement of each data set. The eligibility and the best fit parameter are nowhere being measured in the existing technique and moreover giving the turnaround time higher than the proposed technique. The classical technique is generating and classifying the data items in the cluster that may not be useful in the knowledge discovery.

In the proposed technique, an exclusive measurement is taken into consideration and implemented on the same transaction data for analysis of the results of the proposed as compared to the existing technique. In the proposed scenario, the results obtained are efficient in terms of the generation and inclusion in the clusters as well as the execution time.

Using the exclusive and unique way to measure the eligibility as well as the inclusion of the relevant data items in the best fit cluster based on intelligence, the graph has been plotted to represent the pattern and behavior of the cluster formation process. Using the implementation of cluster formation, it is shown that the proposed technique is producing better results as compared to the classical technique. The graph has been plotted for the warehouse records with respect to the execution time. The investigation has been performed for the slabs or layers of the records for the efficiency analysis.

CONCLUSION AND FUTURE SCOPE

Cluster Formation or simply clustering is the process of aggregation the set of objects in such a manner that objects in the same group called cluster that are more similar in some sense or another to each other than to those in other groups or clusters. It is a prominent and mandatory task of exploratory data mining, and a common technique for statistical data analysis used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, and bioinformatics. Cluster analysis itself is not one specific algorithm, but the general task to be solved. It can be achieved by various algorithms that differ significantly in their notion of what constitutes a cluster and how to efficiently find them. Popular notions of clusters include groups with small distances among the cluster members, dense areas of the data space, intervals or particular statistical distributions. Clustering can therefore be formulated as a multi-objective optimization problem.

A massive amount of research work is under process throughout the globe in assorted algorithms. In this research work, we have proposed and implemented a novel algorithm that makes use of the mathematical foundation and evolutionary approach for the formation of clusters in efficient and effective manners in terms of execution time and associated results. A sample data set of shopping mart has been implemented and the algorithm performs in excellent manner on the desired aspects.

This area of implemented is not limited to the shopping and market survey. The presented and implemented approach can be used in multiple diversified areas including web log usage, forensic investigation, pattern analysis, biometric, astronomy and many other streams that require the efficient methods of aggregation simply called clustering.

The future scope of the research work can extended to the hybrid approach. The hybrid approach makes use of two or more algorithmic approaches to be merged in single formulation to get the optimal results. The hybrid approach can make use of the ant colony optimization or genetic algorithm to get the optimal results. If the presented algorithm is executed to n iterations with genetic algorithmic approach, the best solution can be achieved. In the future work, the cluster formation can be integrated with best first search of the heuristic search methods for the removal of noise.

REFERENCES

[1] Achtert, E.; Böhm, C.; Kröger, P. (2006). "DeLi-Clu: Boosting Robustness, Completeness, Usability, and Efficiency of Hierarchical Clustering by a Closest Pair Ranking". LNCS: Advances in Knowledge Discovery and Data Mining. Lecture Notes in Computer Science 3918: 119–128. doi:10.1007/11731139_16. ISBN 978-3-540-33206-0.

[2] Aditya Desai, Himanshu Singh, Vikram Pudi, 2011. DISC: Data-Intensive Similarity Measure for Categorical Data, Pacific-Asia Conferences on Knowledge Discovery Data Mining

[3] Agrawal, R.; Imieliński, T.; Swami, A. (1993). "Mining association rules between sets of items in large databases". Proceedings of the 1993 ACM SIGMOD international conference on Management of data - SIGMOD '93. p. 207. doi:10.1145/170035.170072. ISBN 0897915925.

[4] Andre Baresel, Harmen Sthamer, Michael Schmidt,2002. Fitness Function Design to improve Evolutionary Structural Testing

[5] Andrew L.Nelson, Gregory J.Barlow, Lefteris Doitsidis,2008 .Fitness Functions in Evolutionary Robotics: A Survey and Analysis

[6] Barnett, V. and Lewis, T.: 1994, Outliers in Statistical Data. John Wiley & Sons., 3rd edition.

[7] Can, F.; Ozkarahan, E. A. (1990). "Concepts and effectiveness of the cover-coefficient-based clustering methodology for text databases". ACM Transactions on Database Systems 15 (4): 483. doi:10.1145/99935.99938

[8] David L. Olsen, Dursun Delen, Advances data mining techniques, Springer, 2008

[9] Fayyad, Usama; Piatetsky-Shapiro, Gregory; Smyth, Padhraic (1996). "From Data Mining to Knowledge Discovery in Databases". Retrieved 17 December 2008

[10] Hans-Peter Kriegel, Peer Kröger, Jörg Sander, Arthur Zimek (2011). "Density-based Clustering". WIREs Data Mining and Knowledge Discovery 1 (3): 231–240. doi:10.1002/widm.30.

[11] He Zengyou, Xu Xiaofei, Deng Shenchun, 2002. Squeezer: An Efficient Algorithm for Clustering Categorical Data,Journal of Computer Science and Technology,Vol. 17, No. 5,pp 611-624

[12] He Zengyou, Xu Xiaofei, Deng Shenchun, 2003. Discovering Cluster Based Local Outliers,Article Published in Journal Pattern Recognition Letters, Volume 24. Issue 9-10,pp 1641-1650,01 June 2003

[13] He Zengyou, Xu Xiaofei, Deng Shenchun, 2006. Improving Categorical Data Clustering Algorithm by Weighting Uncommon Attribute Value Matches, ComSIS Vol.3,No.1

[14] Jerzy Stefanowski, 2009, Data Mining - Clustering, University of Technology, Poland

[15] Lloyd, S. (1982). "Least squares quantization in PCM". IEEE Transactions on Information Theory 28 (2): 129–137. doi:10.1109/TIT.1982.1056489.

[16] M.Davarynejad, M.-R.Akbarzadeh-T, N.Pariz,2007. A Novel Framework for Evolutionary Optimization: Adaptive Fuzzy Fitness Granulation, IEEE Conference on Evolutionary Computation, pp 951-956,2007

[17] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu (1996). "A density-based algorithm for discovering clusters in large spatial databases with noise". In Evangelos Simoudis, Jiawei Han, Usama M. Fayyad. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96). AAAI Press. pp. 226–231. ISBN 1-57735-004-9.

[18] Max Bramer, Principles of data mining, Springer, 2007

[19] Microsoft academic search: most cited data mining articles: DBSCAN is on rank 24, when accessed on: 4/18/2010

[20] Mihael Ankerst, Markus M. Breunig, Hans-Peter Kriegel, Jörg Sander (1999). "OPTICS: Ordering Points To Identify the Clustering Structure". ACM SIGMOD international conference on Management of data. ACM Press. pp. 49–60.

[21] "Outlier Detection in Clustering" ftp://cs.joensuu.fi/pub/Theses/2005_MSc_Cherednichenko_Svethlena.pdf

[22] R. Ng and J. Han. "Efficient and effective clustering method for spatial data mining". In: Proceedings of the 20th VLDB Conference, pages 144-155, Santiago, Chile, 1994.

[23] R.Ranjini, S.Anitha Elavarasi, J.Akilandeswari.2012. Categorical Data Clustering Using Cosine Based Similarity for Enhancing the Accuracy of Squeezer Algorithm

[24] S Roy, D K Bhattacharyya (2005). "An Approach to find Embedded Clusters Using Density Based Techniques". LNCS Vol.3816. Springer Verlag. pp. 523–535.

[25] Shyam Boriah, Varun Chandola, Vipin Kumar, 2008. Similarity Measures for Categorical Data: A Comparative Evaluation, SIAM International Conference on Data Mining-SDM

[26] Tian Zhang, Raghu Ramakrishnan, Miron Livny. "An Efficient Data Clustering Method for Very Large Databases." In: Proc. Int'l Conf. on Management of Data, ACM SIGMOD, pp. 103–114.

[27] Varun Chandola, Arindam Banerjee, Vipin Kumar. Outlier Detection: A Survey

[28] Z. Huang. "Extensions to the k-means algorithm for clustering large data sets with categorical values". Data Mining and Knowledge Discovery, 2:283–304, 1998.