International Journal of Computing and Corporate Research
ISSN (Online) : 2249-054X
Volume 4 Issue 3 May 2014
International Manuscript ID : 2249054XV4I3052014-23

THE PRAGMATIC APPRAISE ON DYNAMIC CLUSTERING ALGORITHM APPROACHES

Amit Chaudhary
Assistant Professor, CSE-IT DEPT
Modern Institute of Engineering and Technology, Mohri
Kurukshetra, Haryana, India

Heena Goyal
M.Tech. (CSE)
Modern Institute of Engineering and Technology, Mohri
Kurukshetra, Haryana, India

Abstract

Data mining refers to the investigation of the huge quantities of data sets stored in computers. Masses of information produced from money registers, from examining, from subject particular databases all around the organization, are investigated, examined, lessened, and reused. Quests are performed crosswise over diverse models proposed for anticipating deals, promoting reaction, and benefit. Traditional factual methodologies are principal to information mining. Mechanized AI strategies are additionally utilized. Information mining obliges ID of an issue, alongside accumulation of information that can prompt better understanding and machine models to give factual or different method for investigation. Information comes in, perhaps from numerous sources. It is incorporated and put in some normal information store. A piece of it is then taken and preprocessed into a standard organization. This 'arranged information' is then moved to an information mining calculation which handles a yield as standards or some other sort of patterns. Clustering or categorization is mandatory in every knowledge discovery in databases (KDD) applications. It is the approach of aggregation of a set of physical objects into classes of analogous objects or homogenous behavior. Cluster formation also generates the pattern and rules that lead to the inclusion of the outlier nodes or data sets. An outlier is appears to deviate markedly from the other associates of the sample where it occurs. In this manuscript, we have analyzed the clustering algorithms for multiple applications.

Keywords - Data Mining, Clustering, Dynamic Clustering, Outlier

International Journal of Computing and Corporate Research
ISSN (Online) : 2249-054X
Volume 4 Issue 3 May 2014
International Manuscript ID : 2249054XV4I3052014-23

INTRODUCTION

Numerous analytic computer models have been used in the domain of data mining. The standard model types in data mining include normal regression for prediction, logistic regression for classification, neural networks, and decision trees. These techniques are well known in the academic and research domains.

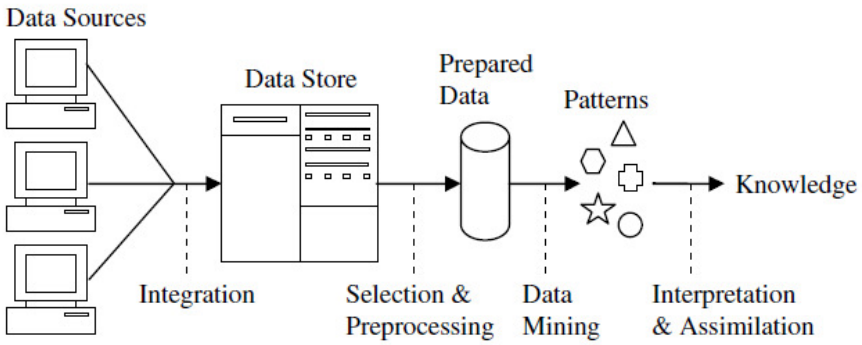


Figure 1.1: Data Mining and Knowledge Discovery Process

Data mining needs the identification of a problem, with the collection of the data that can lead to better understanding and computer models to deliver statistical or other means of analysis. This may be aided by visualization situated apparatuses, which show information, through major measurable examination, for example, relationship dissection. Information mining apparatuses need to be flexible, adaptable, fit for correctly anticipating reactions between movements and outcomes, and equipped for programmed usage. Flexible alludes to the capability of the device to be connected in a wide mixture of models. Versatile instruments allude that if the apparatuses takes a shot at a little information set, it ought to additionally chip away at bigger information sets. Mechanization is valuable, yet its provision is relative. Some diagnostic capacities are frequently robotized, yet human setup before actualizing strategies is needed. Truth be told, examiner judgment is discriminating to fruitful usage of information mining. Legitimate determination of information to incorporate in inquiries is discriminating. Information conversion additionally is regularly needed. An excess of variables transform an excessive amount of yield, while excessively few can disregard key connections in the information. Essential understanding of factual ideas is compulsory for effective information mining.

Information mining [9] alludes to the investigation of the expansive amounts of information that are put away in machines. Information mining has been called exploratory information investigation, in addition to everything else. Masses of information produced from money registers, from checking, from subject particular databases all around the organization, are investigated, broke down, decreased,

International Journal of Computing and Corporate Research
ISSN (Online) : 2249-054X
Volume 4 Issue 3 May 2014
International Manuscript ID : 2249054XV4I3052014-23

and reused. Information mining obliges ID of an issue, alongside gathering of information that can prompt better understanding and workstation models to give measurable or different method for dissection [8].

Clustering is an important KDD technique with numerous applications, such as marketing and customer segmentation. Clustering typically groups data into the sets in such way that the intra-cluster analogous behavior or similarity is maximized and while inter-cluster similarity is minimized [11]. 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].

NOTION OF CLUSTERING AND RELATED ASPECTS

Clustering is the important knowledge discovery technique with numerous applications, such as marketing and customer segmentation. Clustering group data into sets in such a way that the intra-cluster similarity is maximized and while inter-cluster similarity is minimized. Clustering is the form of unsupervised learning that 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.

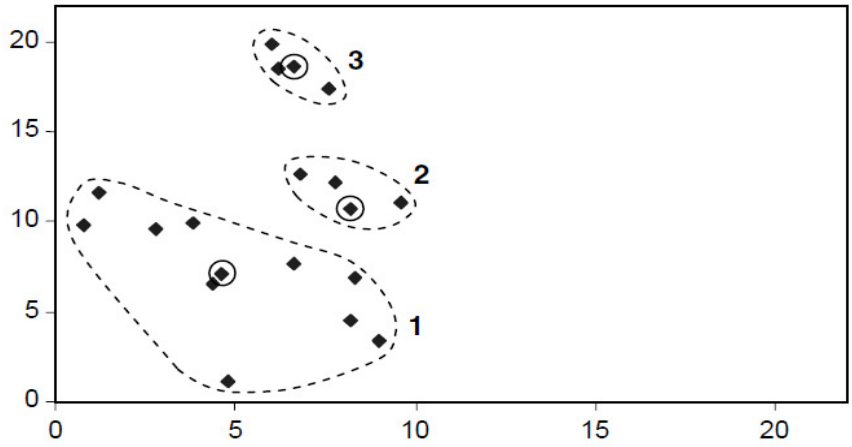


Figure 1.2: Clustering of Data

A lot of people prior bunching calculations concentrate on the numerical information whose characteristic geometric properties might be abused regularly to characterize separation works between

International Journal of Computing and Corporate Research
ISSN (Online) : 2249-054X
Volume 4 Issue 3 May 2014
International Manuscript ID : 2249054XV4I3052014-23

information focuses. Then again, a great part of the information existed in the databases is downright, where quality qualities can't be characteristically requested as numerical qualities. Because of the extraordinary properties of all out characteristics, the grouping of absolute information appears to be more confused than that of numerical information. To beat this issue, a few information-driven similitude measures have been proposed for absolute information. The conduct of such measures straightforwardly relies on upon the information.

Clustering is the most paramount unsupervised taking in issue; in this way, as every other issue of this kind, it manages discovering a structure in an accumulation of unlabeled information. A detached meaning of grouping could be "the methodology of arranging articles into gatherings whose parts are comparable somehow". A group is hence an accumulation of articles which are "comparative" between them and are "different" to the items fitting in with different groups.

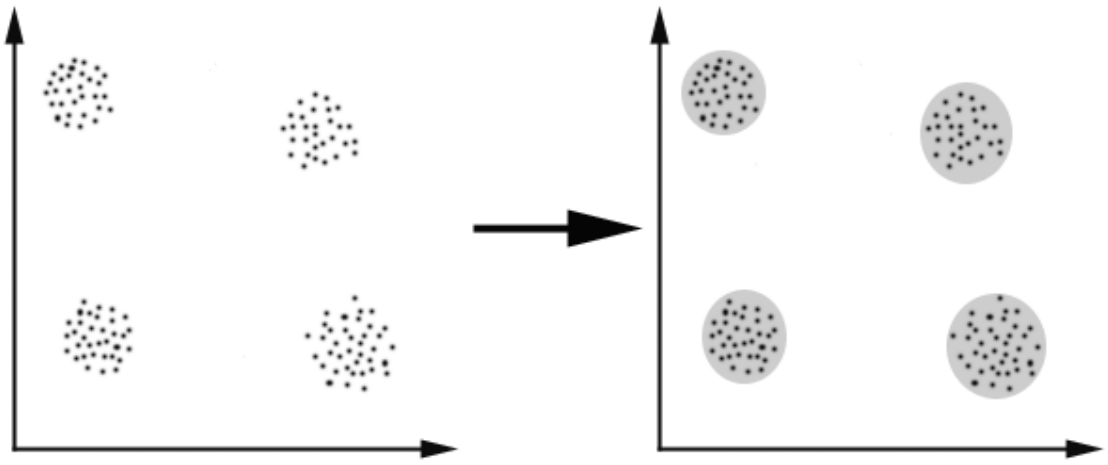


Figure 1.3: Formation of Clusters

Here, we effortlessly distinguish the 4 groups into which the information might be separated; the comparability paradigm is separation: two or more protests fit in with the same group in the event that they are "close" as indicated by a given separation (for this situation geometrical distance).this is called separation-based grouping.

International Journal of Computing and Corporate Research
ISSN (Online) : 2249-054X
Volume 4 Issue 3 May 2014
International Manuscript ID : 2249054XV4I3052014-23

An alternate sort of grouping is calculated bunching: two or more protests have a place with the same group if this one characterizes an idea regular to all that questions. As such, questions are assembled as indicated by their fit to unmistakable ideas, not as per basic likeness measures.

Significant FOCUS IN CLUSTERING

The objective of grouping is to focus the natural gathering 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 must supply this rule, in such a route, to the point that the consequence of the bunching will suit their needs. Case in point, we could be intrigued by discovering delegates for homogeneous gatherings (information decrease), in discovering "common bunches" and portray their obscure properties ("characteristic" information sorts), in discovering helpful and suitable groupings ("valuable" information classes) or in discovering uncommon information objects (outlier recognition).

RELATED WORK

Zengyou He et al proposed Squeezer calculation, a bunching calculation for straight out information. It takes n tuples as info and produces groups as yield. At first, the first tuple is perused and group structure is built. Read consequent tuples one after an alternate. For every tuple, figure its similitudes with all current groups. Select the biggest comparability esteem. On the off chance that the biggest similitude quality is more excellent than limit 's', the tuple is embedded into the current group else new bunch is framed. The Cluster Structure (CS) will be upgraded for every cycle. Squeezer calculation makes utilization of Cluster Structure which comprises of group data and rundown data [4].

André Baresel et al proposed Evolutionary Structural Testing. It utilizes Evolutionary Algorithms (EA) to hunt down particular test information that give high structural scope of the product under test. A vital normal for evolutionary structural testing is that the wellness capacity is built on the premise of the product under test. The wellness capacity itself is not of enthusiasm for the issue; be that as it may, a decently-built wellness capacity might considerably expand the shot of discovering an answer and arriving at higher scope. Better direction of the hunt can bring about advancements with less emphasess, accordingly prompting investment funds in asset consumption. This paper presents exploration comes about on proposed adjustments to the wellness capacity prompting the change of evolutionary testability by accomplishing higher scope with less assets. A set of issues and their particular results are examined [2]

Zengyou He et al proposed FindCBLOF Algorithm for detecting outliers. This algorithm computes the value of CBLOF for each record which determines the degree of record's deviation. This algorithm is efficient for handling large datasets [5].

Zengyou He et al proposed NabSqueezer algorithm, an improved Squeezer algorithm. NabSqueezer algorithm gives more weight to uncommon attribute value matches for finding similarity in similarity computation of Squeezer algorithm. In this algorithm weight of each attribute is precalculated using More Similar Attribute Value Set (MSFVS) method [6].

M. Davarynejad et al proposed computational many-sided quality which is a significant test in evolutionary calculations because of their requirement for rehashed wellness capacity assessments. Here, we intend to diminish number of wellness capacity assessments by the utilization of wellness granulation through a versatile fluffy likeness examination. In the proposed calculation, an unique's wellness is just figured in the event that it has deficient similitude to a queue of fluffy granules whose wellness has recently been registered. In the event that a distinct is sufficiently like a known fluffy granule, then that granule's wellness is utilized rather as a rough gauge. Generally, that distinct is added to the queue as another fluffy granule. The queue measure and every granule's span of impact is versatile and will develop/psychologist relying upon the populace wellness and the amount of disparate granules. The proposed strategy is connected to a situated of 6 conventional enhancement benchmarks that are for their different aspects. In correlation with standard requisition of evolutionary calculations, factual examination uncovers that the proposed system will altogether diminish the amount of wellness capacity assessments while discovering just as great or better results [7].

Shyam Boriah et al introduced a relative study on number of similitude measures, for example, Goodall, Occurence Frequency, Overlap, Inverse Occurence Frequency, Burnbay, Gambaryan, Smirnov. In this paper we have examined the execution of a mixture of closeness measures in the setting of a particular information mining undertaking: outlier discovery [9].

Andrew L. Nelson et al highlights the investigation of wellness capacities utilized within the stream and space of mechanical autonomy. This space is the stream of research that applies manufactured advancement to concentrate the control frameworks for independent robots. In this original copy and examination work, robots endeavor to execute the assignment in a given environment utilizing wellness capacity. The controllers in the upgraded performing robots are chosen, corrupted and engendered to execute the assignment again in an iterative movement that imitates a few parts of common

International Journal of Computing and Corporate Research
ISSN (Online) : 2249-054X
Volume 4 Issue 3 May 2014
International Manuscript ID : 2249054XV4I3052014-23

development. A key segment of this process one may contend, the key part is the estimation of wellness in the advancing controllers. ER is one of a group of machine taking in strategies that depend on communication with, and sentiment from, a complex element environment to drive blend of controllers for self-sufficient executors. These systems can possibly prompt the advancement of robots that can adjust to uncharacterized situations and which may have the capacity to perform undertakings that human fashioners don't totally get it. Keeping in mind the end goal to attain this, issues with respect to wellness assessment must be tended to. In this paper we study ebb and flow ER research and concentrate on work that included genuine robots. The studied exploration is sorted out as per the level of from the earlier information used to detail the different wellness capacities utilized throughout development. The underlying inspiration for this is to distinguish routines that permit the improvement of the best level of novel control, while obliging the base measure of from the earlier errand learning from the architect [3].

Aditya Desai et al introduced closeness which are neighborhood-based or fuse the similitude processing into the taking in calculation. These measures process the area of an information point however not suitable for figuring likeness between a couple of information examples X and Y [1].

R.ranjani et al proposed Enhanced Squeezer calculation, which fuses Data-Intensive Similarity Measure for Categorical Data (DISK) in Squeezer Algorithm. Circle measure, group information by understanding area of the dataset, in this way bunches structured are not simply focused around recurrence dissemination as numerous closeness measures do [8].

TAXONOMY

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 representativescj, as to minimize the mean squared Euclidean distance from each data object to its nearest centroid.

Hierarchical clustering

International Journal of Computing and Corporate Research
ISSN (Online) : 2249-054X
Volume 4 Issue 3 May 2014
International Manuscript ID : 2249054XV4I3052014-23

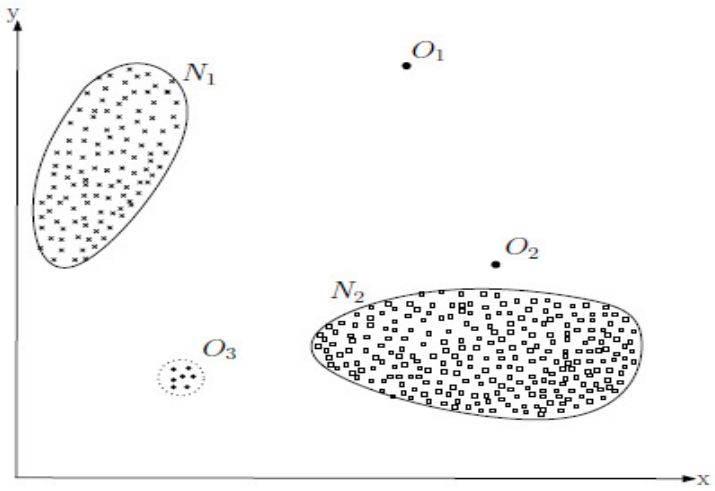
Hierarchical clustering methods build a cluster hierarchy, i.e. a tree of clusters also known as dendogram. A dendrogram is a tree diagram often used to represent the results of a cluster analysis. Hierarchical clustering methods are categorized into agglomerative (bottom-up) and divisive (top-down) 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.

Density-based and grid-based clustering

The key idea of density-based methods is that for each object of a cluster the neighbourhood 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 densitybasedcategory.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.

Outliers

An outlying observation, or outlier[24], 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 Computing and Corporate Research
ISSN (Online) : 2249-054X
Volume 4 Issue 3 May 2014
International Manuscript ID : 2249054XV4I3052014-23

Figure1.4: Outliers in two-dimensional dataset

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 statisticalmethods 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 neighbourhood. 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 dmindistance 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 xjbe 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.

International Journal of Computing and Corporate Research
ISSN (Online) : 2249-054X
Volume 4 Issue 3 May 2014
International Manuscript ID : 2249054XV4I3052014-23

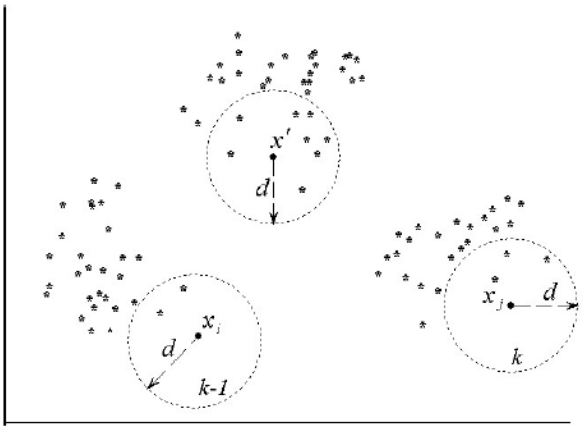


Figure 1.5: Illustration of outlier definition by Knorr and Ng.

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.

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 neighborhood 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.

RELATED ASPECTS OF THE CLUSTER FORMATION PROBLEM

Given a set of unclassified training data sets

- To find an efficient way of partitioning and classifying the training data into classes.
- To construct the representation that enables the category of cluster of any new example to be determined.
- Although the two subtasks are logically distinct, they are usually performed together.

- Classification learning programs are successful if the predictions they make are correct. i.e. If they agree with an externally defined classification.

In clustering, there is no externally defined notion of correctness.

There are a huge number of ways in which a training set could be partitioned.

Some of these are better than others.

The classical methods suggest members of a cluster should resemble each other more than resemble members of other classes.

Hence a good partition should

- Maximise similarity within classes
- Minimise similarity between classes.

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-clusters 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.

International Journal of Computing and Corporate Research
ISSN (Online) : 2249-054X
Volume 4 Issue 3 May 2014
International Manuscript ID : 2249054XV4I3052014-23

	Size of Dataset	Number of Clusters	Type of Dataset	Type of Software
k-means Alg.	Huge Dataset & Small Dataset	Large number of clusters & Small number of clusters	Ideal Dataset & Random Dataset	LNKnet Package & Cluster and TreeView Package
HC Alg.	Huge Dataset & Small Dataset	Large number of clusters & Small number of clusters	Ideal Dataset & Random Dataset	LNKnet Package & Cluster and TreeView Package
SOM Alg.	Huge Dataset & Small Dataset	Large number of clusters & Small number of clusters	Ideal Dataset & Random Dataset	LNKnet Package & Cluster and TreeView Package
EM Alg.	Huge Dataset & Small Dataset	Large number of clusters & Small number of clusters	Ideal Dataset & Random Dataset	LNKnet Package & Cluster and TreeView Package

Table 1 : Comparative Analysis of assorted clustering algorithms

Conclusion

Cluster Formation or basically grouping is the procedure of total the set of articles in such a way, to the point that protests in the same assembly called group that are more comparable in some sense or an alternate to one another than to those in different aggregations or Clusters. It is an unmistakable and required errand of exploratory information mining, and a basic system for factual information examination utilized as a part of numerous fields, including machine taking in, example distinguishment, picture dissection, data recovery, and bioinformatics. Group investigation itself is not one particular calculation, however the general assignment to be explained. It could be attained by different calculations that vary altogether in their idea of what constitutes a group and how to productively discover them. Famous thoughts of groups incorporate aggregations with little separations around the Cluster parts, thick ranges of the information space, interims or specific measurable disseminations. Clustering can hence be figured as a multi-objective enhancement issue. A huge measure of examination work is under methodology all around the globe in different calculations. In this examination work, we have proposed and executed a novel calculation that makes utilization of the

numerical establishment and evolutionary methodology for the shaping of Clusters in productive and successful behavior regarding execution time and cohorted outcomes. An example information set of shopping store has been actualized and the calculation performs in incredible way on the wanted perspectives. What's to come extent of the examination work can stretched out to the cross breed methodology. The mixture methodology makes utilization of two or more algorithmic methodologies to be consolidated in single plan to get the ideal effects. The mixture methodology can make utilization of the ground dwelling insect state enhancement or hereditary calculation to get the ideal outcomes. In the event that the displayed calculation is executed to the emphases with hereditary algorithmic methodology, the best result might be attained. Later on work, the bunch structuring could be coordinated with best first pursuit of the heuristic quest strategies for the evacuation of clamor.

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, VikramPudi, 2011. DISC: Data-Intensive Similarity Measure for Categorical Data, Pacific-Asia Conferences on Knowledge Discovery Data Mining

[3] Andre Baresel, HarmenSthamer, Michael Schmidt,2002. Fitness Function Design to improve Evolutionary Structural Testing

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

[5] Can, F.; Ozkarahan, E. A. (1990). "Concepts and effectiveness of the cover-coefficient-based clustering methodology for text databases".ACM Transactions on Database Systems15 (4): 483. doi:10.1145/99935.99938

[6] Hans-Peter Kriegel, Peer Kröger, Jörg Sander, Arthur Zimek (2011). "Density-based Clustering".WIREs Data Mining and Knowledge Discovery1 (3): 231–240. doi:10.1002/widm.30.

[7] 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

[8] 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

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

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

[11] Lloyd, S. (1982). "Least squares quantization in PCM". IEEE Transactions on Information Theory28 (2): 129–137. doi:10.1109/TIT.1982.1056489.

[12] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu (1996). "A density-based algorithm for discovering clusters in large spatial databases with noise". In EvangelosSimoudis, 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.

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

[14] 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

[15] MihaelAnkerst, 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.

[16] 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.

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

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

[19] ShyamBoriah, VarunChandola, Vipin Kumar, 2008. Similarity Measures for Categorical Data: A Comparative Evaluation, SIAM International Conference on Data Mining-SDM

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

[21] 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.

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

[23] 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.

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