SURVEY ON TRADITIONAL AND EVOLUTIONARY CLUSTERING APPROACHES

Christabel Williams[1], Bright Gee Varghese R[2]

[1]PG Student, [2]Assistant professor, Department of Computer Science and Engineering, Karunya University, Tamil Nadu, India, ***christabel.williams99@gmail.com, brightfsona@yahoo.co.in***

Abstract

Clustering deals with grouping up of similar objects. Unlike classification, clustering tries to group a set of objects and find whether there is some relationship between the objects whereas in classification a set of predefined classes will be known and it is enough to find which class a object belongs. Simply, classification is a supervised learning technique and clustering is an unsupervised learning technique. Clustering techniques apply when there is no class to be predicted but rather when the instances are to be divided into natural groups. These clusters presumably reflect some mechanism at work in the domain from which instances are drawn, a mechanism that causes some instances to bear a stronger resemblance to each other than they do to the remaining instances. It naturally requires different techniques to the classification and association learning methods .Clustering has many applications in various fields. In Software engineering it helps in reverse engineering, software maintenance and for re-building systems. It aims at breaking a larger problem into small pieces of understanding elements. There are many approaches available to carry out clustering. Since clustering has no particular methodology there are many methods available for carrying out clustering. There are many traditional as well as evolutionary methods available for carrying out clustering. In this paper various types of the above mentioned methods are described and some of them are compared. Each method has its own advantage and they can be used according to the needs of the user.

Keywords: *Clustering, Classification, Software Engineering, Traditional, Evolutionary.*

----------***----------

1. INTRODUCTION

Clustering is a technique for finding similarity groups in data, called clusters. It groups data instances that are similar to (near) each other in one cluster and data instances that are very different (far away) from each other into different clusters. Clustering is often called an unsupervised learning task as no class values denoting an a priori grouping of the data instances are given, which is the case in supervised learning. Clustering can be used in software maintenance and re-use. In order to maintain software, understanding the source code is mandatory. Therefore source code should be clustered. The main purpose is to understand the system , artifacts recovery and identifying the relationships among the source code.

Clustering can be done in many ways and there are several methods for carrying out clustering .Traditional methods includes partitional based, hierarchical and density based clustering. There are some evolutionary clustering approaches as well like relocation approaches, grid based and density based [1].

The partitional approaches includes k-means, graph-theoretic clustering and density based clustering and so on. These approaches construct various partitions and evaluate them using some criterion. On the other hand evolutionary approaches are used for solving optimization problems. Here the evolutionary operators and clustering structures are converged to give an optimal solution. Both these approaches are useful according to the circumstances in which they are used. Each of them have their own pros and cons and can be used effectively according to the user needs.

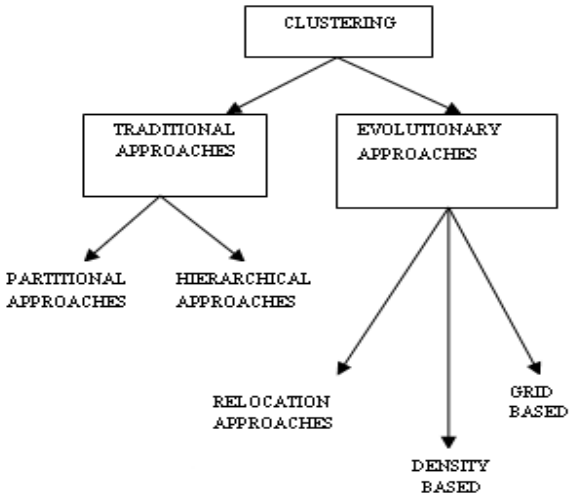

Fig.1 Various Clustering approaches

2. VARIOUS CLUSTERING APPROACHES

This section focuses on the two approaches of clustering: Traditional and evolutionary approaches.

2.1 Traditional Approaches

The traditional approaches of clustering include three categories–partitional, hierarchical and density based approaches. Since clustering do not have any precise notion there are many clustering methods available to carry out clustering. According to [2] Farley and Raftery (1998) clustering can be divided into two main groups – hierarchical and partitioning methods. Apart from these several methods can proposed based on the induction principle.

2.1.1 Hierarchical Methods

In hierarchical clustering clusters are built by recursively partitioning the instances in either a top down or a bottom-down fashion.

Agglomerative Hierarchical Clustering: It is a bottom-up approach in which each object itself is a cluster of its own. Then these clusters are merged until a desired cluster structure is obtained. Ying Zhao and George Karypis [3] defines certain clustering criterion that can be used to determine which clusters to merge at each step.

[3] For example in an *n* document dataset and the solution has been obtained after *m* merging steps. The solution includes *n-1* clusters as one cluster, as one cluster will be removed in each step. So accordingly the next pair to be merged will be selected which will lead to a *n-2* solution that optimizes the selected clustering criterion. That is one of the (n-1)*(n-2)/2 pairs of merges will be evaluated and the particular clustering criterion with the minimum value is selected.the criterion function will be locally optimized in that particular stage of the algorithm. Since agglomerative algorithm has computationally expensive steps and its complexity cannot be reduced for any particular clustering criterion it is not effective.

Divisive Hierarchical Clustering: [22] Christopher D. Manning, Prabhakar Raghavan and Hinnich Schutze says that top-down clustering is conceptually more complex than bottom-up clustering since a second, flat clustering algorithm as a subroutine is needed.. Divisive algorithm has the advantage of being more efficient if a complete hierarchy is not generated down to individual document leaves. For a fixed number of top levels, using an efficient flat algorithm like *k*-means, top-down algorithms are linear in the number of documents and clusters. So they run much faster than HAC algorithms. Bottom-up methods make clustering decisions based on local patterns without initially taking into account the global distribution. These early decisions cannot be undone. Top-down clustering benefits from complete information about the global distribution when making top-level partitioning decisions

The result of a hierarchical clustering is a dendogram including the nested grouping of objects with the similarity level. The clustering can be obtained by cutting the dendogram at the desired similarity level. [9] Lior Rokach and Oded Maimon.

[4] Jain et al 1999 classifies the hierarchical clustering further based on the manner in which their similarity measures are calculated.

Single-linkage Clustering: Also known as minimum method or nearest neighbour method .In this method the distance between the two clusters are said to be equal to the shortest distance from any member of one cluster to any member of other cluster. If similarities are present in the data then the similarity between two clusters will be equal to the greatest similarity from any member of one cluster to any member of other cluster. [5] Sneath and Sokal 1973. [6] Guha et al 1998 summarizes the disadvantages as chaining effect. A few points forming a bridge between two clusters may cause the two clusters to become one.

Complete-link Clustering: Also known as maximum method or furthest neighbour method. Here the distance between two clusters are considered to be equal to the longest distance from member belonging to one cluster to other member belonging to another cluster [7] King 1967.

Average-link Clustering: Minimum variance method. Here the distance between two clusters are said to be equal to the average distance from any member of one cluster to any member of another cluster. But the disadvantage is that it may cause the elongated clusters to split and neighboring clusters to merge.

The main disadvantage of hierarchical clustering is these methods can never undo what has been done previously ie., it cannot be backtracked. And since its complexity is more it requires huge i/o costs.

2.1.2 Partitioning Methods

Partitioning methods of clustering are flexible methods based on the iterative relocation of data points between clusters. The quality is measured by a clustering criterion [8] Sami A¨ yra¨mo¨ Tommi Ka¨rkka¨inen

[21] Witten, Ian H, Eibe Frank,2005,says that clustering techniques apply when there is no class to be predicted but rather when the instances are to be divided into natural groups. Clustering naturally requires different techniques to the classification and association learning methods. The classic clustering technique is called k-means. First, the number of

clusters needed should be mentioned in a parameter called k. Then k points are chosen at random as cluster centers. According to the Euclidean distance metric every instance is assigned to its closest cluster center. Next the mean, of the instances in each cluster is calculated—this is the "means" part. These means forms the new center value of the corresponding clusters. The whole process is repeated with the new cluster centers. Iteration continues until same points are assigned to each cluster in all rounds, and this indicates the cluster centers are stabilized and will remain the same forever. This clustering method is simple and effective. It is easy to prove that choosing the cluster center to be the center minimizes the total squared distance from each of the cluster's points to its center. Once the iteration has over, each point is assigned to its nearest cluster center, so the overall effect is to minimize the total squared distance from all points to their cluster centers. But the minimum is a local one; there is no guarantee that it is the global minimum. If there is a change in the initial random choice fully different arrangements can arise. It is almost always infeasible to find globally optimal clusters.

2.2 Evolutionary Approaches

Clustering lacks in a general objective and this is the main reason for having many clustering methods. Nature has been offering us many concepts or ideas for solving optimization problems faced by modern human society. Evolutionary computation techniques are general methods for solving optimization problems and since clustering can be considered as an optimization problem it can be solved using evolutionary approaches.

[9] Lior rokach and oded maimon et al says that evolutionary operators and a clustering population can be used to converged into a global optimal solution. Clustering components are used as chromosomes and evolutionary operators are selection, recombination and mutation. In EC fitness function value is calculated and those chromosomes that are having good fitness values are able to survive to the next level. Genetic algorithm (GA) is the most frequently used evolutionary algorithm. Each cluster will be associated with a particular fitness value. Since fitness value is inversely proportional to the squared error, clusters with small squared value will have high fitness value.

In GA's selection operator promotes solution from one generation to the next level depending on the fitness value. Selection is carried out based on the probabilistic values depending on the fitness value. Crossover is the most popularly used recombination operator in use. It takes a pair of chromosomes as inputs and produces a pair of offspring. Mutation can also be used. But a major problem with GA is their sensitivity to various parameters like population size, crossover and mutation etc. Many researchers suggested various guidelines but of no use. Better results can be obtained by using hybrid genetic algorithms.

2.2.1 Relocation Approaches

Since in EC techniques solutions evolve from one generation to other in an iterative process , clustering is carried out using relocation methods. At first the idea of using evolutionary techniques to carry out clustering was proposed by [10] Krovi et al in 1991 where a genetic algorithm was proposed to split the data into two clusters. Here the solution are strings of integers of length that of size of the dataset. Comparatively this algorithm suffers from drawbacks like redundancy and invalidity. The algorithm maximizes the ratio between the sum of squares and within the sum of squares. This was used by [11] Krishna and Narasimha murty1999 for searching a partition with a fixed number of clusters using modified genetic operators. Instead of crossover single k-means iteration is used.

Partition is carried out as a boolean k x n matrix as proposed by [12] Bezdek et al 1994 and focuses on reducing the sum-of-squared-errors criterion. Experiments are carried out with different distance metrics in order to detect clusters of different shapes. Here the crossover operator simply swaps the columns between chromosomes and mutation changes the cluster assignment of one object randomly.

[13] Luchian et al 1994 came up with a new encoding in which search for simultaneous optimum number of clusters and optimum partition are allowed. Partition is carried out similar to k-means , according to the proximities data items are assigned to clusters. Crossover and mutation operators are extended to work with variable-length chromosomes and real encoding. A new operator called Lamarckian operator acting at gene level which modifies a cluster to match the mean of the corresponding cluster is introduced. This could be called as hybridization with k-means because the cluster assignment procedure and the updating of the cluster representative leads to one iteration if the traditional clustering. But this leads to premature convergence due to an increased selection pressure. [14] Hall et al(1999) extended the same algorithm for carrying out searching in fuzzy partitions with fixed number of clusters. Cluster elements are represented using gray coding and this became the most successful clustering literature [15]Maulik and Bandyopadhyay 2000. This kind of approaches are based on the distances between the data items and the cluster centres.

[13] Luchian et al 1994 proposed a centroid based encoding, which gives continuous optimization. They are mostly used in clustering based on PSO and are also useful to search for cluster representatives. Their performance is comparatively higher when compared to that of k –means algorithm and are much more better than or equal to k –means provided with best initial configuration. This is due to the increased exploration capabilities and are not much strongly dependent

on initialization. [16] Abraham et al 2007 presented a survey on swarm intelligence techniques. Centroid based encoding and differential evolution was used in supervised as well as unsupervised scenario.

2.2.2 Density Based Approaches:

In density-based approaches for clustering, multi-modal evolutionary algorithms that search for cluster centres lying in the dense region in the feature space are used [17] Dumitrrescu and Simon 2003. The correctness of the cluster centroids are measured using the Gaussian functions. [18] Nasraoui et al 2005 says that when data are distributed according to the normal distributions then each one of the clusters will be a hyper-ellipsoid associated with a mean and a covariance matrix. Local maxima of a density function are traced out using a multi-modal genetic algorithm. Each individual is represented using a point in the m-dimensional space in order to identify centres of the dense regions.

[19] Zaharie et al 2005 observes the use of a differential evolution algorithm in the same context. It is not only concerned with the number of clusters but also on the hyper-ellipsoid scales. The method proposed by [18] Nasraoui et al 2005 generates only one descriptor whereas in this approach asset of descriptors can be associated to the same cluster. It provides a reliable identification of clusters in noisy data.

2.2.3 Grid Based Approaches:

[20] Sarafis et al 2002 proposed a genetic algorithm that searches for a partition of the feature space that also provides a partition of the data set. The algorithm generates rules that will build grid in the space. Each individual has a set of k clustering rules corresponding to one cluster. In turn each rule is derived from m genes and each gene corresponds to interval involving one feature. Many attempts were made to minimize square-error criterion by using a flexible fitness function and it has a high computational cost because of the form of fitness function.

3. CONCLUSIONS

This paper gives a survey on various techniques for carrying out clustering. It covered both traditional as well as evolutionary approaches. Evolutionary approaches are performing better than traditional approaches. But each approaches have their own performance levels.[3] Ying Zhao and George Karypis says that for every criterion function, partitional algorithms always lead to better clustering results than agglomerative algorithms, which suggests that partitional clustering algorithms are well-suited for clustering large document datasets due to not only their relatively low computational requirements, but also comparable or even better clustering performance.This review takes into account many algorithms and reviewed them according to their performance.

ACKNOWLEDGEMENTS

I am highly indebted to Mr. Bright Gee Varghese, Assistant Professor, Department of Computer Science and Engineering for his guidance and I would like to thank all the authors of the references which were very helpful for this survey.

REFERENCES

[1]. Clustering: Evolutionary Approaches Mihaela Elena Breaban ,Prof. PhD Henri Luchian
[2]. Fraley C. and Raftery A.E., “How Many Clusters? Which Clustering Method? Answers Via Model-Based Cluster Analysis”, Technical Report No. 329. Department of Statistics University of Washington, 1998
[3]. Comparison of Agglomerative and Partitional Document Clustering Algorithms- Ying Zhao and George Karypis. Department of Computer Science, University of Minnesota, Minneapolis, MN 55455
[4]. Jain, A.K. Murty, M.N. and Flynn, P.J. Data Clustering: A Survey. ACM Computing Surveys, Vol. 31, No. 3, September 1999
[5]. Sneath, P., and Sokal, R. Numerical Taxonomy. W.H. Freeman Co., San Francisco, CA, 1973
[6]. Guha, S., Rastogi, R. and Shim, K. CURE: An efficient clustering algorithm for large databases. In Proceedings of ACM SIGMOD International Conference on Management of Data, pages 73-84, New York, 1998.
[7]. King, B. Step-wise Clustering Procedures, J. Am. Stat. Assoc. 69, pp. 86-101, 1967.
[8]. “Introduction to partitioning based clustering methods with a robust example” Sami A¨ yra¨mo¨ Tommi Ka¨rkka¨inen
[9]. “Clustering Methods” Lior Rokach Department of Industrial Engineering Tel-Aviv University Oded Maimon Department of Industrial Engineering Tel-Aviv University.
[10]. Ravindra Krovi. Genetic algorithms for. clustering: A preliminary investigation. In Proceedings of the Twenty-Fifth Hawaii International Conference on System Sciences, pages 540{544. IEEE Computer Society Press, 1991.
[11]. K. Krishna and M. Narasimha Murty. Genetic k-means algorithm. Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on, 29(3):433 {439, June 1999. ISSN 1083-4419. doi: 10.1109/3477.764879.
[12]. James C. Bezdek, Srinivas Boggavarapu, Lawrence O. Hall, and Amine Bensaid. Genetic algorithm guided clustering. In International Conference on Evolutionary Computation, pages 34{39, 1994
[13]. Silvia Luchian, Henri Luchian, and Mihai Petriuc. Evolutionary automated classification. In Proceedings of 1st Congress on Evolutionary Computation, pages 585{588, 1994
[14]. L. O. Hall, B. Ozyurt, and J. C. Bezdek. Clustering with a genetically optimized approach. IEEE Transactions on Evolutionary Computation, 3:103{112, 1999
[15]. Ujjwal Maulik and Sanghamitra Bandyopadhyay. Genetic algorithm-based clustering technique Pattern Recognition, 33(9):1455 { 1465, 2000. ISSN 0031-3203

[16]. Ajith Abraham, Swagatam Das, and Sandip Roy. Swarm intelligence algorithms for data clustering. Soft Computing for Knowledge Discovery and Data Mining, Springer Verlag, pages 279{313, 2007.
[17]. D. Dumitrescu and Kroly Simon. Evolutionary prototype selection. In Proceedings of the International Conference on Theory and Applications of Mathematics and Informatics ICTAMI, pages 183{190, 2003
[18]. Olfa Nasraoui, Elizabeth Leon, and Raghu Krishnapuram. Unsupervised niche clustering: Discovering an unknown number of clusters in noisy data sets. In Ashish Ghosh and Lakhmi Jain, editors, Evolutionary Computation in Data Mining, volume 163 of Studies in Fuzziness and Soft Computing, pages 157{188. Springer Berlin Heidelberg, 2005.
[19]. Daniela Zaharie. Density based clustering with crowding differential evolution. Symbolic and Numeric Algorithms for Scientific Computing, International Symposium on, pages 343{350, 2005
[20]. I. Sarafis, A. M. S. Zalzala, and P. W. Trinder. A genetic rule-based data clustering toolkit. In Proceedings of the Evolutionary Computation on 2002. CEC '02.Proceedings of the 2002 Congress - Volume 02, CEC '02, pages 1238{1243, Washington, DC, USA, 2002. IEEE Computer Society. ISBN 0-7803-7282-4. URL http://portal.acm.org/citation.cfm?id=1251972.1252396
[21]. Witten, Ian H, Eibe Frank,2005, Data Mining - Practical Machine Learning Tools and Technique" 2nd Edition,Morhan Kaufmann,San Francisco
[22]. Christopher D. Manning, Prabhakar Raghavan and Hinnich Schutze "Introduction to Information Retrieval"

BIOGRAPHIES

Christabel Williams is pursuing MTech (Computer Science and Engineering) degree from Karunya University, Tamil Nadu, India. She received her BE (Computer Science and Engineering) degree from C.S.I College of Engineering, Ketti, The Nilgiris. Tamil Nadu, India

Bright Gee Varghese. R completed his BE (CSE) in Sun College of Engineering and Technology, Nagercoil, Tamil Nadu, India in 2003. He started his career as Lecturer in Sun College of Engineering and Technology, Nagercoil, from June 2003. He completed his M.E from Vinayaka Mission University, Salem. Currently he is working as an Assistant Professor in Computer Science Department in Karunya University and pursuing part time Ph.D in Karunya University Tamil Nadu, India. His main research area is Software Engineering.