An Intelligent Fuzzy Convex Hull based Clustering Approach

Rita Keshari
M.Tech (CSE) Scholar
ABES Engineering College
Ghaziabad, India

Amit Sinha
Associate Professor
ABES Engineering College
Ghaziabad, India

ABSTRACT
Data mining refers to the extraction of knowledge by analyzing the data from different perspectives and accumulates them to form an useful information which could help the decision makers to take appropriate decisions. Classification and clustering has been the two broad areas in data mining. As the classification is a supervised learning approach, the clustering is an unsupervised learning approach and hence can be performed without the supervision of the domain experts. The basic concept is to group the objects in such a way so that the similar objects are closer to each. In this paper, an approach is made by fusing the concept of convex hull with fuzziness parameter. Each boundary data point is validated for the convex hull property to form a cluster. The decision value depends on the membership value of the particular point. The points satisfying the convex hull property forms a cluster. The performance

General Terms
Data mining, clustering

Keywords
Data mining, Clustering, Convexity, Tangent, Convex hull.

1. INTRODUCTION
Data mining has become an important area of research due to the availability of huge amount of data and the necessity of extracting the meaningful information from this data [1]. It also focuses on analyzing the relationship among the data and finds the hidden patterns in the data. The different data mining tools such as classification, clustering, association rule mining, etc has been useful for decision making.

Clustering is the process of grouping the data objects together on the basis of their similar characteristics [2]. The data objects are dissimilar to the objects of a group are placed in other groups. A measure is used for finding such similarity among the objects. Clustering is done to reduce the huge amount of data into smaller manageable units so that the extraction of knowledge can be done easily.

A number of clustering algorithms have been proposed which are somehow better than the one another. The existing algorithms have different categories on the basis of the type of measure and the technique formulated for cluster formation. Among these most of the algorithms [1],[2],[3] such as K-means, K-medoid, CLARA, CLARANS, etc. are dependent on the basic factor of stating the number of clusters *k* in the data set [9] . Therefore the application of such clustering algorithms is not prominent in the areas where the number of clusters in not known. However, the algorithms such as DBSCAN [4], OPTICS [9], CURE [5], etc. try to solve the problem of *k.* Although these algorithms have tried to address the problem of *k,* but the determination of some other parameters which would result in optimal clusters in a data set is a difficult task.

In this paper we have tried to look into the clustering approach in some different manner. Here the concept of convex hull [8] is introduced with the fuzziness assigned to each object in order to form a cluster. Each data point is validated with the property of convex hull property [9] to in order to decide the boundary of the cluster.

2. LITERATURE REVIEW
A number of research works has been done in the field of data mining for finding the best clustering algorithm to group the data. Among enormous approaches some of the approaches have been detailed here. The *K*-means [6] clustering algorithm is a simple partitioning approach. Initially the number of cluster is defined in this algorithm. The centroid of cluster is calculated in each iteration by finding the distance of each data point to each cluster center. The closest data point is placed to the respective cluster. This procedure is repeated until all the data points are in proper groups and stabilize themselves. However, this procedure does not yield good result when the initial center points are not chosen properly.

K-medoid [6] is another partitioning technique of clustering which divides the data sets into *K* number of clusters. The initial centroid value is determined by finding the medoid value. Each data point is associated with the closet medoid. In each iteration, the position of medoid is re-calculated and the algorithm terminates when the medoids become fixed. The *K*-medoid tries to find a non-overlapping set of clusters.

CLARA (Clustering LARge Applications) [2] is a clustering algorithm to handle large data sets. This algorithm basically relies on sampling technique. A sample from the data set is obtained on which the PAM (Partitioning Around Medoids) algorithm is applied to find the medoids. A sample is drawn randomly yielding the medoids results in approximating the medoids for the entire data set. The clustering is done by calculating average dissimilarity for the data point. If the value is less then the current minimum value, the updation for the centroid is done else the medoids are retained. The algorithms iterate until the medoids are stabilized. CLARA draws many samples to find the medoid which would result in better clustering.

The hierarchical clustering for a data set is done in two ways: Agglomerative clustering and Divisive clustering [6].

Agglomerative clustering is done by calculating the proximity between the two clusters. The algorithm starts with individual clusters and merges closest pair of clusters on the basis of their proximity values. Divisive clustering starts with the

entire set of clusters and splits into singleton cluster. A dendogram, an inverted tree describes the order in which the clusters are merged and spitted in either method.

DBSCAN [4] is a density based clustering algorithm. The algorithm has *minpts* and *epsilon* input parameters instead of specifying the number of clusters *k*. For obtaining the exact clusters the tuning of these parameters is required. The data points are categorised as the core points and border points. The core points lie in the density of the cluster whereas the border points have less number of points in their neighbourhood and are incapable to form clusters. The basic idea is to find the density around a point by evaluating the neighbourhood density reachability and density connectivity of the points. Those points which lie in a denser region are considered to be cluster points. However, tuning of the *minpts* and *epsilon* parameter is a tedious job for finding the exact clusters.

3. PROPOSED WORK

3.1 Problem Formulation

Most of the existing clustering algorithms focus on the similarity measure for the purpose of clustering which becomes difficult when large data set is considered for the analysis. Moreover the exact boundary of the clusters is not precisely defined. Therefore, the convex hull property [15] is used for finding the similar data points to be clustered. Moreover, when the boundary around a cluster of points can be drawn the points have low inter-cluster value and high intra-cluster value. Hence the vagueness of point for the belongingness to a particular cluster can be removed.

3.2 Convex Hull

A convex set is a set where the line segment joining two points from the same set lies within the set. In a convex hull [15][16] the region is bounded by a cycle of line segments joining one end to the other forming a closed region. The lines joining to each other are known as edges. For a plane the smallest convex hull is the region which has the convexity within that subset. The basic property for the convex hull formation is to analyze the turn of the line segment when the two points from the set are joined together. If the intersection of the points results in a left turn for the next line segment, the line is connected from the initial to the final point resulting in right turn of the line segment.

3.3 Methodology

The proposed methodology includes the introduction of the fuzziness parameter in convex hull for the decision making of the boundary points in the convex hull. The existing clustering method basically deals with the similarity measure for cluster formation. Here we have tried to look the clustering method in some different aspect. The cluster points can be viewed as the convex points lying within the data set. A convex hull encapsulates the data points which are convex in nature [11]. A boundary is drawn around the similar data points on the basis of their fuzziness value. This fuzzy value is determined by a membership assignment of each data point [14]. The degree of belongingness gives the grouping of the data point. The cluster formation is done after the convex hull around the set of given points is constructed.

3.3.1 Initialization of membership

The data point is selected random and the dissimilarity measure for the points in the data set is calculated. A fuzzy membership value [13] is calculated for each data point.

$$mem(i) = \frac{p_i - p_{\max}}{p_{\max} - p_{\min}}$$

.....Equation 1

$$mem(p_i) = 1 - mem(i)$$

.....Equation 2

3.3.2 Convex hull property

The data points are sorted according to the membership values and lowest, first lowest and second lowest values are considered for each iteration. The feature value of each selected data point is obtained from the input data. The convex hull property as the cross product is calculated as given in equation (3).

Let p_0, p_1, p_2 be three vector points with feature values $a_0, b_0, c_0, d_0, a_1, b_1, c_1, d_1 and \quad a_2, b_2, c_2, d_2$ respectively.

Therefore tangent formula for convex hull is given as:

$$(p_1 - p_0)\mathrm{X}(p_2 - p_0) = (a_1 - a_0)(b_2 - b_0)(c_1 - c_0)(d_2 - d_0) - (a_2 - a_0)(b_1 - b_0)(c_2 - c_0)(d_1 - d_0)$$

.....Equation 3

3.3.3 Formation of cluster

Each data points obtained after step 3.3.2 are checked for the property of cluster formation. The data points' having the counter-clockwise direction tends to form cluster. The data points with next lowest membership value are updated accordingly. A cluster is formed when difference between membership values of two consecutive data points is greater than the specified value α. The set of data points are removed from the data set and the algorithm is iterated for next cluster formation.

3.3.4 Termination condition

The algorithm terminates when all the data points gets exhausted.

3.4 Algorithm Steps

The algorithm has been divided into three different phases.

Algorithm

Input: Set of data points D,
Threshold value α.

Output: Set of clusters C.

Phase 1:

q= select an initial random data point from D;

for i : 1 to n

find the distance of q to each point in D.

$$dist(p,q) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2}$$

end

for i : 1 to n

Assign membership to each data point

$$mem(i) = \frac{p_i - p_{\max}}{p_{\max} - p_{\min}}$$

$$mem(p_i) = 1 - mem(i)$$

end

Phase 2:

Do

for i : 1 to n

Set the points for convex hull formation

$$lowest = mem(p_i)$$

$$lowest1 = mem(p_i + 1)$$

$$lowest2 = mem(p_i + 2)$$

Get the feature values ($\dot{f}$) of the lowest, lowest1 and lowest2.

Calculate the tangent on ($\dot{f}$) using cross product of the feature values.

Check for the convex hull property by identifying counter-clockwise direction.

Add to C_i data point p_i.

Update the next set of points for convex hull formation

$$lowest = mem(p_i)$$

$$lowest1 = mem(p_i + 1)$$

$$lowest2 = mem(p_i + 2)$$

end

Phase 3:

if $diff\left(mem(p_i), mem(p_i + 1)\right) > \alpha$

Remove the set of points C_i from D.

end

Until $D \neq \phi$.

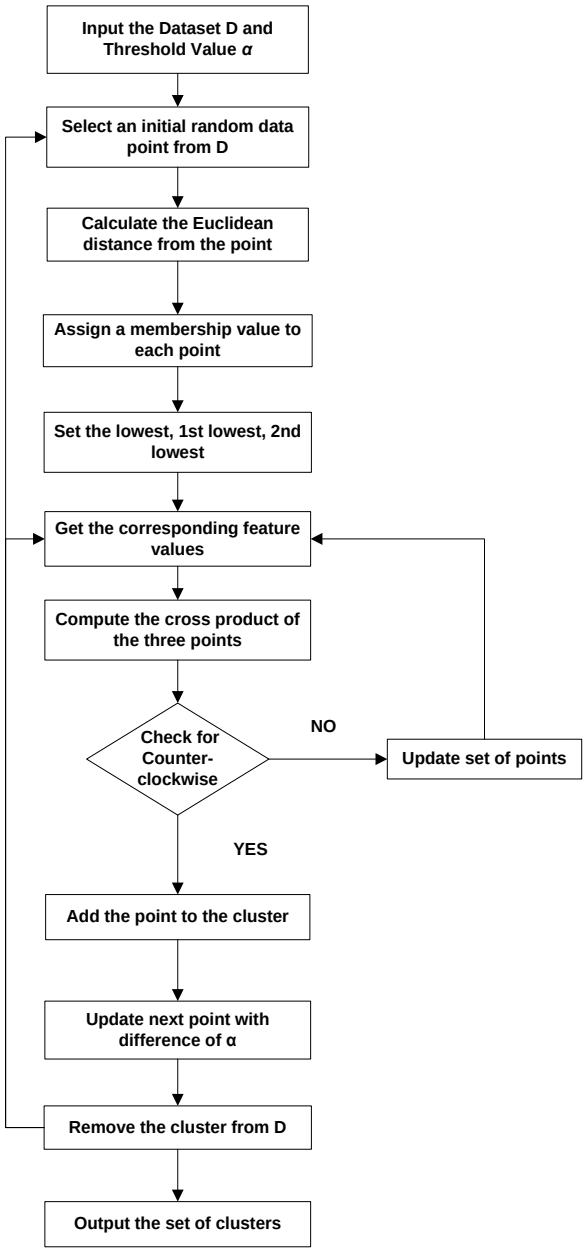

Figure 1 : Flowchart of proposed algorithm

The proposed algorithm tries to capture the central idea of clustering and providing an optimal set of clusters such that the overlapped of cluster points can be easily identified by defining the boundary around the points. Therefore, the vagueness of data points at boundary towards the belongingness to a particular cluster becomes comparatively quite less.

4. RESULT

The experiment is performed on a 2.93 Gigahertz Intel Core i3 processor computer with 3 GB memory, running on Windows 2007. The algorithm is implemented in C programming language.

Figure 2 Graph of clustering accuracy

In this experiment the IRIS Flower data set from UCI repository having 150 records and 4 feature values with three clusters is considered. From experimental analysis the following observation is derived. In the PAM clustering algorithm, the selection of the medoids is done randomly at the beginning. The medoid values are calculated in the successive iterations. The accuracy of the algorithm is 67% as the medoids for perform the clustering by combining two clusters based on their similarity measure. The above figure shows the CLARA clustering algorithm has 89.34% of accuracy as the medoid selection is done on sampling basis with the PAM clustering algorithm. Therefore, the algorithm finds its medoids from the sample more accurately. In the proposed approach, as the convex hull formation for each cluster is done thereby it reduces the chance of misclassification resulting in 96.67% of accuracy. The clusters obtained have drawn a boundary around its data points in order to get more optimum set of clusters.

5. CONCLUSION AND FUTURE WORK

Clustering has a number of applications in every field of life. The work presented gives a detailed description of the existing clustering methods highlighting the deficiency of the methods. The proposed algorithm tries to address the problem stating number of clusters and parameter adjustment. The application of convex hull for clustering results in cluster formation with boundary around the data points of that cluster. The comparative analysis of the proposed algorithm gives better result the existing methods.

In the broader area of application this algorithm can enhanced with possibilistic approach of clustering being with the fuzziness parameter of the algorithm

6. ACKNOWLEDGMENTS

The support and technical assistance provided at the computer labs of ABES Engineering College, Ghaziabad INDIA, is highly appreciated without which the research of this paper could not have been completed. We also sincerely thanks the faculty members of CSE department, ABES Engineering College for the valuable guidance provided which helped us in this research through various stages.

7. REFERENCES

[1] Ravichandran,I. 2003, Data mining and clustering techniques, Technical Report.

[2] Jain, A. and Dubes, R. 1988. Algorithms for Clustering Data. Prentice-Hall, Englewood Cliffs, NJ.

[3] Ertoz, L., Stienbach, M., and Kumar, V. 2002. Finding clusters of different sizes, shapes, and densities in noisy, high dimensional data, Technical Report.

[4] Ester, M., Kriegel, H-P., Sander, J. and XU, X. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In Proceedings of the 2nd ACM SIGKDD, 226-231, Portland, Oregon.

[5] Guha, S., Rastogi, R., and Shim, K. 1998. CURE: An efficient clustering algorithm for large databases. In Proceedings of the ACM SIGMOD Conference, 73-84, Seattle, WA.

[6] Han, J. and Kamber, M. 2001. Data Mining. Morgan Kaufmann Publishers.

[7] J.Wang, B. Yang , W. Zhang and B. Qin, "Convex hull-based support vector machine rule extraction", 9th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), page 689 – 692, 2012.

[8] A. A. Ramli, J. Watada and W. Pedrycz, "Real Time Model of Fuzzy Random Regression Based on a Convex Hull Approach", International Conference on Advances in Computing, Control, and Telecommunication Technologies, pages 45-49, 2010.

[9] S. Theodoridis and K. Koutroumbas, "Pattern Recognition", fourth edition,Elsevier, 2009.

[10] H. Liu, S. Xiong ; Q. Chen, "Fuzzy Support Vector Machines Based on Convex Hulls", IEEE International Symposium on Knowledge Acquisition and Modeling, pages 920-923, 2008.

[11] W. Pedeyez, J. V. De, and Oliveria, "Advances in Fuzzy Clustering and its Applications", John Wiley & Sons, New York, 2007.

[12] P. D´ıaz, D. R. Llanos, B. Palop, "Parallelizing 2D-Convex Hulls on clusters: Sorting matters", XV Journal on Parallelism, 2004.

[13] S. Nascimento, B. Mirkin, and F. Moura Pires, "Modeling proportional membership in fuzzy clustering", IEEE Transactions on Fuzzy Systems, 11(2):173–186, 2003.

[14] S. Nascimento, B. Mirkin, and F. Moura Pires, "A fuzzy clustering model of data fuzzy c- means", In the Ninth IEEE International Conference on Fuzzy Systems,volume 1, pages 302–307, 2000.

[15] C. B. Barber, D. P. Dobkin and H. HuhdanPaa, "The Quickhull Algorithm for ConvexHulls", ACM Transactions on Mathematical Software, 22 (4), 1996.

[16] B.B. Chaudhari, "Fuzzy convex hull determination in 2-D space", Pattern Recognition Letters, volume 12(10), pages 591-594, 1991.