2008 International Conference on Advanced Computer Theory and Engineering

Fuzzy Clustering Methods in Data Mining: A comparative Case Analysis

Raju G[1], Binu Thomas[2], Sonam Tobgay[3] and Th. Shanta Kumar[4]
[1]SCMS School of Technology & Management, India. kurupgraju@rediffmail.com
[2] Research Scholar, Mahatma Gandhi University,Kerala, India.
[3] Sherubtse College, Kanglung, Bhutan.
[4]Research Scholar, Himachal Pradesh University,India

Abstract

The conventional clustering algorithms in data mining like k-means algorithm have difficulties in handling the challenges posed by the collection of natural data which is often vague and uncertain. The modeling of imprecise and qualitative knowledge, as well as handling of uncertainty at various stages is possible through the use of fuzzy sets. Fuzzy logic is capable of supporting to a reasonable extent, human type reasoning in natural form by allowing partial membership for data items in fuzzy subsets. Integration of fuzzy logic in data mining has become a powerful tool in handling natural data. In this paper we introduce the concept of fuzzy clustering and also the benefits of incorporating fuzzy logic in data mining. Finally this paper provides a comparative analysis of two fuzzy clustering algorithms namely fuzzy c-means algorithm and adaptive fuzzy clustering algorithm.

1. Introduction

Data Mining or Knowledge discovery refers to a variety of techniques that have developed in the fields of databases, machine learning and pattern recognition [1]. The process of finding useful patterns and information from raw data is often known as Knowledge discovery in databases or KDD. Data mining is a particular step in this process involving the application of specific algorithms for extracting patterns (models) from data [5].Cluster analysis is a technique for breaking data down into related components in such a way that patterns and order becomes visible. It aims at sifting through large volumes of data in order to reveal useful information in the form of new relationships, patterns, or clusters, for decision-making by a user. Clusters are natural groupings of data items based on similarity metrics or probability density models. Cluster analysis has the virtue of strengthening the exposure of patterns and behavior as more and more data becomes available [7]. A cluster has a center of gravity which is basically the weighted average of the cluster. Membership of a data item in a cluster can be determined by measuring the distance from each cluster center to the data point [6].

This paper provides an overview of the crisp clustering technique, advantages and limitations of fuzzy c-means clustering in comparison with adaptive fuzzy clustering method which is superior to c-means clustering in handling outlier points. Section 2 describes the basic notions of clustering and also introduces k-means clustering algorithm. In Section 3 we explain the concept of vagueness and uncertainty in natural data. Section 4 introduces the fuzzy c-means clustering algorithm and describes how it can handle vagueness and uncertainty. Section 5 is about the adaptive fuzzy clustering method which is superior to the c-means algorithm. Section 6 demonstrates the presented concepts in the paper. Finally, section 7 concludes the paper.

2. Crisp clustering techniques

Traditional clustering techniques attempt to segment data by grouping related attributes in uniquely defined clusters. Each data point in the sample space is assigned to only one cluster. In partitioning the data only cluster centers are moved and none of the data points are moved [12]. Thus clustering is an iterative process of finding better and better cluster centers in each step. K-means algorithm and its different variations are the most well-known and commonly used partitioning methods. The value 'k' stands for the number of cluster seeds initially provided for the algorithm. This algorithm takes the input parameter 'k' and partitions a set of m objects into k clusters [7]. The technique work by computing the distance between a data point and the cluster center to add an item into one of the clusters so that intra-cluster similarity is high but inter-cluster similarity is low. A common method to find the distance is to calculate to sum of

978-0-7695-3489-3/08 $25.00 © 2008 IEEE
DOI 10.1109/ICACTE.2008.199

the squared difference as follows and it is known as the Euclidian distance [10](exp.1).

$$d_k = \sum_{n} \left\| X_j^k - Ci_j \right\|^2 \qquad (1)$$

where,

d_k : is the distance of the k^{th} data point

n : is the number of attributes in a cluster

X_j^k : is j^{th} value of the k^{th} data point

C_j^i : is the j^{th} value of the i^{th} cluster center

The cluster centers are randomly initialized and we assign a data point x_i into a cluster to which it has minimum distance. When all the data points have been assigned to clusters, new cluster centers are calculated by finding the weighted average of all data points in a cluster. The cluster center calculation causes the previous centroid location to move towards the center of the cluster set. This is continued until there is no change in cluster centers. The k means algorithm is efficient in handling crisp data where we have clear cut boundaries.

3. Uncertainty and vagueness

The real world data is almost never arranged in such clear cut groups. Instead, clusters have ill defined boundaries that smear into the data space often overlapping the perimeters of surrounding clusters [4]. This happens because natural data do not happen to appear in clear-cut crisp fashion but suffers from the following limitations:

1) Not clearly known: Questionable; problematical
2) Vague : Not definite of determined
3) Doubtful : Not having certain information
4) Ambiguous : Many interpretations
5) Not steady : Varying
6) Liable to change : Not dependable or reliable

When we further examine these meanings two categories of limitations emerge quite naturally – Uncertainty and Vagueness [8]. Vagueness is associated with the difficulty of making sharp and precise distinctions in the world. Uncertainty is a situation in which the choice between two or more alternatives is left unspecified [11]. The modeling of imprecise and qualitative knowledge, as well as handling of uncertainty at various stages is possible through the use of fuzzy sets. Fuzzy logic is capable of supporting, to a reasonable extent, human type reasoning in natural form by allowing partial membership for data items in fuzzy subsets [2].

4. Fuzzy clustering algorithm

The central idea in fuzzy clustering is the non-unique partitioning of the data in a collection of clusters. The data points are assigned membership values for each of the clusters. The fuzzy clustering algorithms allow the clusters to grow into their natural shapes [15]. In some cases the membership value may be zero indicating that the data point is not a member of the cluster under consideration. Many crisp clustering techniques have difficulties in handling extreme outliers but fuzzy clustering algorithms tend to give them very small membership degree in surrounding clusters [14]. The non-zero membership values, with a maximum of one, show the degree to which the data point represents a cluster. Thus fuzzy clustering provides a flexible and robust method for handling natural data with vagueness and uncertainty.

4.1 Fuzzy c-means algorithm

Fuzzy c-means clustering involves two processes: the calculation of cluster centers and the assignment of points to these centers using a form of Euclidian distance. This process is repeated until the cluster centers stabilize. The algorithm is similar to k-means clustering in many ways but it assigns a membership value to the data items for the clusters within a range of 0 to 1. So it incorporates fuzzy set's concepts of partial membership and forms overlapping clusters to support it. The algorithm needs a fuzzification parameter m in the range [1,n] which determines the degree of fuzziness in the clusters. When m reaches the value of 1 the algorithm works like a crisp partitioning algorithm and for larger values of m the overlapping of clusters is tend to be more. The algorithm calculates the membership value μ with the formula,

$$\mu_j(x_i) = \frac{\left(\frac{1}{d_{ji}}\right)^{\frac{1}{m-1}}}{\sum_{k=1}^{p}\left(\frac{1}{d_{ki}}\right)^{\frac{1}{m-1}}} \qquad (2)$$

where

$\mu_j(x_i)$: is the membership of x_i in the j^{th} cluster

d_{ji} : is the distance of x_i in cluster c_j

m : is the fuzzification parameter

p : is the number of specified clusters

d_{ki} : is the distance of xi in cluster C_k

The new cluster centers are calculated with these membership values using the exp. 4.

$$c_j = \frac{\sum_i \left[\mu_j(x_i)\right]^m x_i}{\sum_i \left[\mu_j(x_i)\right]^m} \qquad (3)$$

where

C_j : is the center of the j^{th} cluster
x_i : is the i^{th} data point
μ_j : the function which returns the membership
m : is the fuzzification parameter

This is a special form of weighted average. We modify the degree of fuzzyness in x_i's current membership and multiply this by x_i. The product obtained is divided by the sum of the fuzzified membership. This way new centroids are calculated for clusters.

Pseudo code of fuzzy c-means clustering algorithm [10] is given below:

initialize p=number of clusters
initialize m=fuzzification parameter
initialize C_j (cluster centers)
Repeat
For i=1 to n:Update $\mu_j(x_i)$ applying(3)
For j=1 to p:Update C_i with(4)with current $\mu_j(x_i)$
Until C_j estimate stabilize

The first loop of the algorithm calculates membership values for the data points in clusters and the second loop recalculates the cluster centers using these membership values. When the cluster center stabilizes (when there is no change) the algorithm ends.

4.2 Limitations of the algorithm

The fuzzy c-means approach to clustering suffers from several constrains that affect the performance [10]. The main drawbacks are due to the restriction that the sum of membership values of a data point x_i in all the clusters must be equal to one as in expression (4). This restriction tends to give high membership values for the outlier points. So the algorithm has difficulty in handling outlier points. Secondly the membership of a data point in a cluster depends directly on the membership values of other cluster centers and this sometimes happens to produce undesirable results.

$$\sum_{j=1}^{p} \mu_j(x_i) = 1 \qquad (4)$$

5. The adaptive fuzzy clustering method

Adaptive fuzzy clustering algorithm is similar to c-means algorithm in many ways and it supports the concept of partial memberships for data points in clusters. The main difference is that it removes the restrictions imposed in c-means algorithm through expression (4). The algorithm calculates fuzzy membership values for a data points through a new method as given in exp. 5

$$\mu_j(x_i) = \frac{n * \left(\frac{1}{d_{ji}}\right)^{\frac{1}{m-1}}}{\sum_{k=1}^{p}\sum_{z=1}^{n}\left(\frac{1}{d_{kz}}\right)^{\frac{1}{m-1}}} \qquad (5)$$

where

$\mu_j(x_i)$: is the membership of x_i in the j^{th} cluster
d_{ji} : is the distance of x_i in cluster c_j
m : is the fuzzification parameter
p : is the number of specified clusters
n : is the number of data points
d_{ki} : is the distance of xi in cluster C_k

In the place of exp.(4),the algorithm imposes a new constrain given in exp.(6) which says the sum of membership values of all the points in all the cluster centers must be equal to the number of data points n.

$$\sum_{j=1}^{p}\sum_{i=1}^{n} \mu_j(x_i) = n \qquad (6)$$

During the iteration cycle, the algorithm calculates new cluster centers using the same exp.(3).

The maximum membership value generated by adaptive fuzzy clustering algorithm is not limited to one. When conventional fuzzy membership distributions are required (within a range of zero to one) these can be generated by the process of normalization. Normalization finds the maximum membership among all clusters and rescales the memberships from this maximum using exp.(6).

$$\mu_{ik}^{norm}(x_i) = \frac{\mu_{ik}^{old}(x_i)}{\max \mu_k^{old}} \quad \text{i=1 to n; k=1 to p} \qquad (7)$$

Here,

$m_{ik}^{norm}(x_i)$: is the normalized membership of x_i in the k-th cluster
$\mu_{ik}^{old}(x_i)$: is the old (or original) membership
p : is the number of specified clusters
n : is the number of data points

max() : returns the maximum membership value in the k-th cluster

Pseudo code of adaptive fuzzy clustering algorithm [10] is given below:

initialize p=number of clusters
initialize m=fuzzification parameter
initialize C_j (cluster centers)
repeat
for i=1 to n:Update $\mu_j(x_i)$ applying(5)
for j=1 to p:Update C_i with (3) with current $\mu_j(x_i)$
until C_j estimate stabilize
if fuzzy properties are needed
For i= 1 to n
Normalize $\mu_j(x_i)$ applying (7)
end For
end if

5.1 Advantage of the algorithm

The adaptive fuzzy clustering algorithm is efficient in handling data with outlier points or natural data with uncertainty and vagueness since the algorithm is not restricted by the exp. (4). It accomplishes this by virtue of its unique partial membership features for data items in different clusters so that the clusters grow naturally to reveal hidden patterns. The algorithm tends give only small membership values to outlier points which we demonstrate in the following section.

6. Illustration

Consider the monthly income and expenses of fifteen people form Table1. Assume that at some point of time we have two cluster centers at C1(4,4) and C2(14,8) and we want to consider an outlier point at (2,15) (fig.1). Apparently the point does not belong to any of the clusters and this is a vague and uncertain situation. If we use k-means algorithm, the point will be included to C1 which is against the entire concept of clustering (see Table 1). If we try to handle the situation with fuzzy partial memberships with c-means algorithm, as given in Table 1, the point will be given more membership vales in cluster centers because of the constrain given in exp. 4 (fig 1).

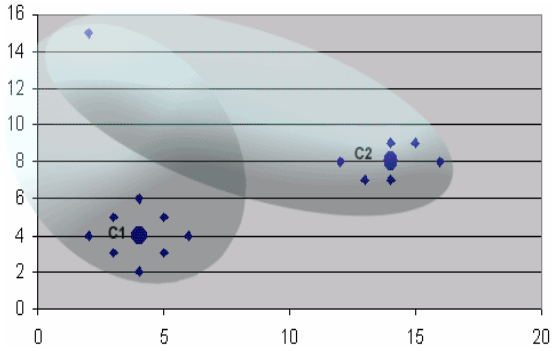

Fig 1. The points and the outcome of c-means algorithm with the outlier point

Table 1. The income(X) and expenses(Y) in 1000s of fifteen people and the inability of both k-means and fuzzy c-means algorithms to handle the outlier point at (2,14)

Points		A1	A2	A3	A4	A5	A6
X	Y						
2	4	2.0	12.6	C1	0.86	0.14	1
3	3	1.4	12.1	C1	0.90	0.10	1
3	5	1.4	11.4	C1	0.89	0.11	1
4	2	2.0	11.7	C1	0.85	0.15	1
4	6	2.0	10.2	C1	0.84	0.16	1
5	3	1.4	10.3	C1	0.88	0.12	1
5	5	1.4	9.5	C1	0.87	0.13	1
6	4	2.0	8.9	C1	0.82	0.18	1
12	8	8.9	2.0	C2	0.18	0.82	1
13	7	9.5	1.4	C2	0.13	0.87	1
14	7	10.4	1.0	C2	0.09	0.91	1
14	9	11.2	1.0	C2	0.08	0.92	1
15	9	12.1	1.4	C2	0.10	0.90	1
16	8	12.6	2.0	C2	0.14	0.86	1
2	15	11.2	13.9	C1?	0.55	0.45	1

Where A1:Distance from C1 (4,4)
A2: Distance from C2 (14,8)
A3: k-means Membership In
A4: Fuzzy C-means Membership in C1
A5: Fuzzy C-means Membership in C2
A6: Sum of fuzzy memberships in C1 &C2 (exp 4)

Table 2. The fuzzy membership values according to the adaptive fuzzy clustering method. The outlier point (2,15) is excluded from both the cluster by assigning very low membership values.

Points								
X	Y	B1	B2	B3	B4	B5	B6	B7
2	4	2	12.65	C1	0.72	0.11	0.71	0.08
3	3	1.4	12.08	C1	1.02	0.12	1.00	0.08
3	5	1.4	11.40	C1	1.02	0.13	1.00	0.09
4	2	2.0	11.66	C1	0.72	0.12	0.71	0.09

4	6	2.0	10.20	C1	0.72	0.14	0.71	0.10
5	3	1.4	10.30	C1	1.02	0.14	1.00	0.10
5	5	1.4	9.49	C1	1.02	0.15	1.00	0.11
6	4	2.0	8.94	C1	0.72	0.16	0.71	0.11
12	8	8.9	2.00	C2	0.16	0.72	0.16	0.50
13	7	9.5	1.41	C2	0.15	1.02	0.15	0.71
14	7	10.4	1.00	C2	0.14	1.44	0.14	1.00
14	9	11.2	1.00	C2	0.13	1.44	0.13	1.00
15	9	12.1	1.41	C2	0.12	1.02	0.12	0.71
16	8	12.6	2.00	C2	0.11	0.72	0.11	0.50
2	15	11.2	13.89	C1	0.13	0.10	0.13	0.07

Where B1: Distance from C1(4,4)
B2: Distance from C2 (14,8)
B3: Membership In
B4: Fuzzy Membership in C1
B5: Fuzzy Membership in C2
B6: Normalized Fuzzy Membership in C1
B7: Normalized Fuzzy Membership in C2

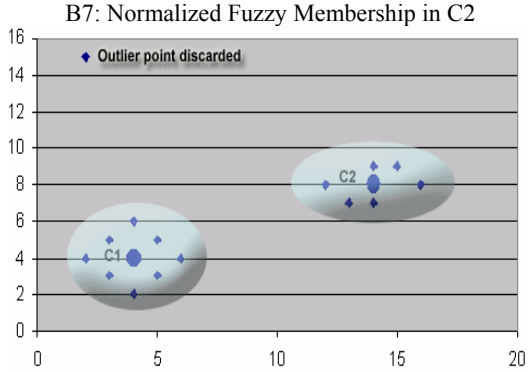

Fig 2. The Outcome of the adaptive fuzzy clustering algorithm (Outlier point is discarded)

7. Conclusion

A good clustering algorithm produces high quality clusters to yield low inter cluster similarity and high intra cluster similarity. Many conventional clustering algorithms like k-means and fuzzy c-means algorithm achieve this on crisp and highly structured data. But they have difficulties in handling unstructured natural data which often contain outlier data points. The adaptive fuzzy clustering algorithm is highly efficient in handling the natural data which suffer from vagueness and uncertainty. The algorithm achieves this by assigning very low membership vales to the outlier points. This fuzzy algorithm is found to be more efficient in obtaining hidden patterns and information from natural data with outlier points.

8. References

[1] K. Pal and P. Mitra, "Data Mining in Soft Computing Framework: A Survey", *IEEE transactions on neural networks, vol. 13, no. 1* (January 2002).

[2] R. Cruse and C. Borgelt, "Fuzzy Data Analysis Challenges and Perspective", http://citeseer.ist.psu.edu/kruse99fuzzy.html

[3] W. H. Au and K.C.C. Chan, "Classification with Degree of Membership: A Fuzzy Approach", *Proceedings IEEE International Conference on Data Mining 2001, ICDM 2001.*

[4] M. Halkidi, "Quality assessment and Uncertainty Handling in Data Mining Process", http://citeseer.ist.psu.edu/halkidi00quality.html

[5] W. H. Inmon, "The data warehouse and data mining" *Commun, ACM, vol. 39,* pp. 49–50 (1996).

[6] U. Fayyad and R. Uthurusamy, "Data mining and knowledge discovery in databases". *Commun, ACM, vol. 39,* pp. 24–27 (1996).

[7] P. Berkhin, "Survey of Clustering Data Mining Techniques", http://citeseer.ist.psu.edu/berkhin02survey.html

[8] Chau, M., Cheng, R., and Kao. B., "Uncertain Data Mining: A New Research Direction". www.business.hku.hk/~mchau/papers/UncertainDataMiningWSA.pdf

[9] Keith C.C, C. Wai-Ho Au, B. Choi, "Mining Fuzzy Rules in A Donor Database for Direct Marketing by A Charitable Organization". *Proceedings First IEEE International Conference on Cognitive Informatics Volume ,* Issue (2002) Page(s): 239 – 246.

[10] E. Cox, "Fuzzy Modeling and Genetic Algorithms for Data Mining and Exploration", Elsevier, (2005).

[11] G. J. Klir and T. A. Folger, "Fuzzy Sets, Uncertainty and Information". Prentice Hall (1988).

[12] J. Han and M. Kamber, "Data Mining Concepts and Techniques". Elsevier (2003).

[13] J. C. Bezdek, "Fuzzy Mathematics in Pattern Classification", *Ph.D. thesis*, Center for Applied Mathematics, Cornell University, Ithica N.Y (1973).

[14] C. G. Looney, "A Fuzzy Clustering and Fuzzy Merging Algorithm", http://citeseer.ist.psu.edu/399498.html

[15] F. Klawonn and A. Keller, "Fuzzy Clustering Based on Modified Distance Measures", http://citeseer.ist.psu.edu/klawonn99fuzzy.html.