International Journal of Applied Engineering Research ISSN 0973-4562 Volume 10, Number 13 (2015)

Relevant Subspace Clustering using Fuzzy Attribute Weighting Approach

[1] Mrs. M. Suguna [2] Dr. S. Palaniammal

[1]Assistant Professor, Department of Computer Science, Dr. GRD College of Science, Coimbatore.
[2]Professor & Head, Department of Science & Humanities, Sri Krishna College of Technology, Coimbatore.
[1] sugunaphd123@gmail.com

Abstract
Recent datasets often contain large number of dimensions which contains clusters only in some specific dimensions that are not known in advance. Clustering algorithm identifies clusters that exist in different subspace. In this paper attribute weight is determined using fuzzy clustering method. The relevant attributes are determined by fuzzy c means method. A FP-tree is constructed for the relevant attributes and top-down strategy is applied to mine dense units from the data. According to the change in density threshold FP-tree is reconstructed. In order to incrementally maintain the dense units bottom-up strategy is applied. After the dense units are mined, the dense units with common faces are merged to form clusters. The clusters are validated using JOSE measure which determines the closeness of objects in the cluster. Experimental results show that the proposed enhanced DENCOS method is mostly effective than the existing techniques in creating superior quality clusters.

Keywords: subspace clustering, weighted fuzzy, FP-tree, JOSE measure

1. Introduction

Clustering is long been a vital research area in varied topic as a significant data processing technique. Data mining and machine learning (Bezdek 1981) are major areas where clustering is used as a basic technique for data analysis and visualization. Though, the crucial task to most traditional clustering algorithms is that in several real world cases, data objects in different clusters are often associated with diverse feature subclasses. For instance, provided a subgroup of data objects recognized in a cluster, it is probable that these objects reveal the similar features as objects from other clusters when a firm subset of dimensions are detected(Jain 2010). Subspace clustering algorithms are proposed during the recent years as an extension to the conventional clustering algorithms in order to handle this issue (Parsons et al 2004). The aim of subspace clustering or projected clustering is to detect clusters with their related dimensions that are embedded in different subspaces of the original data space (Kriegel et al 2009;Muller et al 2009). Depending upon the category in which subspace clusters are determined, they are generally categorized into two main classifications namely hard subspace clustering and soft subspace clustering (Deng et al 2010; Jing et al 2007).

In this paper, soft subspace clustering is taken for study in order to measure the importance of every dimension to a specific cluster in the partitioning process by assigning varied weights to different dimensions. The fuzzy clustering algorithm (Dunn 1973) is modified in order to assign weights to dimensions where the value ranges between 0 and 1. So the dimensions have a partial membership to the subspace which is described in a fuzzy way in the input domain. This new approach, which integrates weighted fuzzy approach, FP-tree construction and validation using deapthness measure and Jose measure, determines the formation of quality clusters. Section 2 presents a brief survey of attribute weight and subspace clustering methods. Fuzzy clustering method is briefed in section 3.Section 4 describes the proposed attributed weighted fuzzy approach with subspace clustering creation. Experimental study on real data is presented in section 5. Finally section 6 concludes the work.

2. Literature Review

Siminski (2012) has proposed a fuzzy subspace clustering approach for high dimensional datasets. The clustering algorithm with weighted attributes minimizes the criterion function. Apart from cluster centers and membership it works upon the weights of descriptors in each cluster. The weights of attributes are assigned with a value between 0 and 1. This shows that the dimensions can have fuzzy membership to the subspace. The method has been used for identification of rule base for fuzzy and neuro-fuzzy systems.

Zhu et al (2014) has presented a scalable clustering technique for streaming data. The soft subspace clustering FWSC-SD uncovers the vital local subspace features of high dimensional data. It is able to handle large scale streaming data. The method partitions the streaming data into chunks and processes every data segment continuously. The size of the chunk varies according to the available memory and speed of the data stream. The weighted cluster centers retain required statistics in scalable clustering to fit with the streaming soft subspace clustering.

AnisurRahman&Zahidul Islam (2012) has propounded a fuzzy clustering technique named CRUDAW which allows a data miner to allocate weights on the attributes of a dataset based on their significance for clustering. Based on the density of the records of a dataset initial seeds are chosen. The initial fuzzy membership degree is also assigned deterministically. The similarity of values is considered to measure the distance between the values of categorical attributes rather than assigning the distance to be either 0 or 1. The work can be extended to automatic generation of attribute weights.

Hongyan Liu &Feng Kong (2005) presented a subjective and objective integrated method that defines attribute weights in fuzzy multiple attribute decision making (FMADM) problems. It determines weights by computing mathematical models automatically and also uses decision maker's preferences in order to overcome the limitation that occur in subjective or objective methods when applied in fuzzy MADM problems.

3. Fuzzy Clustering Method

Fuzzy c-means partitions set of n objects D = {x_1,x_2, ... , x_n} in R^d dimensional space into c (1 < c < n) fuzzy clusters with Z = {z_1, z_2, ... , z_c} cluster centers or centroids. The fuzzy clustering of objects is described by a fuzzy matrix *mp* with n rows and c columns in which n is the number of data objects and c is the number of clusters. mp_{ij}, the element in the i^{th} row and j^{th} column in *mp*, indicates the degree of association or membership function of the i^{th}object with the j^{th} cluster. The characters of *mp* are as follows:

$$mp_{ij} \in [0,1] \ \forall \text{ i= 1,2,....n; } \forall \text{ j=1,2,....c} \quad (1)$$

$$\sum_{i=1}^{n} mp_{ij} = 1 \ , \forall \ j = 1,2, \dots , c \quad (2)$$

$$0 < \sum_{i=1}^{n} mp_{ij} < n \ \forall \text{ j=1,2,....c} \quad (3)$$

The objective function of FCM algorithm is to minimize the Equation (4):

$$J_m = \sum_{i=1}^{c} \quad \sum_{j=1}^{n} mp_{ij}^{m} \, d_{ij}^{2} \ 1 \leq m \ < \alpha \quad (4)$$

Where $d_{ij} = \|x_i - z_j\|$ (5)

in which, m (m > 1) is a scalar termed the weighting exponent and controls the fuzziness of the resulting clusters and d_{ij} is the Euclidean distance from object x_i to the cluster center z_j. The z_j, centroid of the jth cluster, is obtained using Equation(6).

$$Zj = \frac{\sum_{i=1}^{n} \mu_{ij}^{m} x_i}{\sum_{i=1}^{n} \mu_{ij}^{m}} \quad (6)$$

4. Methodology

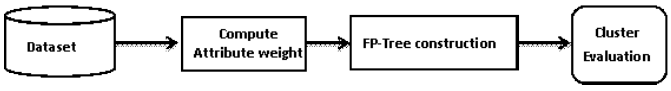

Figure 1. Framework of the proposed method

Figure 1 shows the framework of the model. In this paper an efficient approach for subspace cluster formation is proposed which are evaluated by cluster validity indices.

Step 1: Attribute Weighting Fuzzy Clustering

Attribute weights are determined by applying fuzzy c-means clustering approach. The objective function is to minimize

$$J(X, C, U) = \sum_{i=1}^{n} \sum_{j=1}^{c} \sum_{k=1}^{p} h(u_{ij}) wt_{ik}^{v} (x_{jk} - mp_{ik})^2 \quad (7)$$

where $\sum_{k=1}^{p} m_{ik} = 1 : \sum_{k=1}^{m} wt_{ik} = 1 \ \forall i \ ; 1 \leq i \leq c$

The weights are updated using the equation

$$\forall i \ ; 1 \leq i \leq c : \forall k ; 1 \leq k \leq m : wt_{ik} = \frac{s_{i.k}^{\frac{2}{1-v}}}{\sum_{r=1}^{m} s_{i.r}^{\frac{2}{1-v}}} \quad (8)$$

The attribute with weights are selected for FP-Tree construction which satisfy the threshold *t*.

Step 2: FP-Tree Construction

Attribute weighting is followed by scanning the database to construct the first branch of the tree with the frequent items. Those items that are deleted will be inserted into new database named DB', at the same time the transaction is added to the trailer table and let the new transaction's trail of node link pointing to the last node in that path. If all the items in a transaction are frequent, then the items are added to the FP-tree and that transaction can be deleted directly by not adding any item to the new database DB'. In a transaction if all the items are frequent then no need to add that transaction to the trailer table.

Step 3: Cluster Evaluation

- **Proposed cluster validation measures**

1. Joint Optimum Similarity Eccentric Measure

The attributes of data objects are identified and for each identified attribute the diverse factor is computed which represents the mean value of that particular attribute (Suguna&Palaniammal 2014). Similarly the diversity factor is calculated for each of the attribute identified. Then from the entire diversity factor computed, the overall diversity factor is computed which shows the closeness of data point with the cluster. Similarly this will be performed for each of the cluster and finally a single cluster will be selected which has more closure value. The JOSE measure is computed for both numerical and non-numeric values, in case of non-numeric values, the proposed method uses enum values and rules. The following equation shows the calculation of diversity factory and similarity measure:

$$\text{Diversity factor Df} = \int_{1}^{N} \Sigma(Ai - Aj)/N \text{dx} \quad (14)$$

$$\text{Compute jose measure Jm} = Ø(As)*(\Sigma Df/N) \quad (15)$$

5. Experimental Results

The performance of proposed method is compared with the existing subspace clustering algorithms namely CLIQUE, PROCLUS and DENCOS. The experiments were conducted on breast cancer, heart disease and mushroom dataset. The performances of the algorithm were evaluated based on accuracy and validity (precision, recall, F-measure) measures. The data sets were collected from the UCI data repository. The data sets are varied in terms of size, attributes and classes. Table 1 shows the concise characteristics of the data sets.

Table 1.Datasets

Data sets	Instances	Attributes
Breast cancer	286	9
Mushroom	8124	22
Heart disease	303	75

Table 2 shows the weights of attributes for breast cancer data set. It shows that the attribute clump thickness has a high probability with the value greater 0.6500 which lies in cluster 1(0.7212) and cluster 5(0.8245). The attribute uniformity of cell shape with the value 0.6707 lies in cluster 2. Marginal adhesion lie in cluster 5(0.7991). Bare nuclei lie in cluster 1(0.8726) and cluster 4(0.7031). Normal nucleoli lies in cluster 1(0.7152), cluster 3(0.8158) and cluster 5(0.7721).

Table 2. Weights of attributes for breast cancer dataset

Attribute	Attribute weights in clusters (1 – 5)				
Clump thickness	**0.7212**	0.2478	0.6158	0.6307	**0.8245**
Uniformity of cell size	0.4892	0.4771	0.1056	0.4637	0.3144
Uniformity of cell shape	0.1138	**0.6707**	0.4469	0.3479	0.5371
Marginal adhesion	0.5193	0.2382	0.6233	0.3457	**0.7991**
Single epithelial cell size	0.2142	0.5427	0.2731	0.5933	0.3908
Bare nuclei	**0.8726**	0.0017	0.3821	**0.7031**	0.4371
Bland chromatin	0.4468	0.3638	0.5372	0.4178	0.1559
Normal nucleoli	**0.7152**	0.2925	**0.8158**	0.0027	**0.7721**
Mitoses	0.5263	0.2914	0.5139	0.1246	0.4382

Table 3. Weights of attributes for Heart disease dataset

Attribute	Attribute weights in clusters (1 – 3)		
age	0.3786	0.3634	**0.7390**
sex	**0.7732**	0.4239	**0.6730**
cp	0.4529	**0.8029**	0.2810
trestbps	0.2333	0.4729	0.5390
chol	0.4590	**0.8920**	**0.6792**
fbs	0.3482	**0.6821**	0.4202
restecg	0.2804	0.4021	0.3245
thalach	**0.6592**	**0.7223**	0.5643
exang	**0.6743**	**0.7854**	**0.8765**
oldpeak	0.3489	0.2951	0.3961
slope	0.4783	0.6573	0.4636
ca	0.5918	0.4537	0.3434
thal	0.5346	**0.6567**	**0.8876**

Table 3 shows that the attributes age, sex, cp, chol, fbs, thalach, exang and thal are assigned to three clusters from the subset attributes of heart disease data set. Table 4 shows the attribute weights of mushroom data set. The attributes namely cap-shape, cap-surface, bruises, gill-attachment, stalk-shape, stalk-root, stalk-surface-below-ring, veil-type, veil-color, ring-type, spore-print-color and population are identified from the four different clusters.

Table 4. Weights of attributes for mushroom dataset

Attribute	Attribute weights in clusters (1 – 4)			
cap-shape	0.3217	**0.6782**	**0.7210**	0.5292
cap-surface	0.4382	0.3492	0.2478	**0.8254**
cap-color	0.6013	0.3116	0.6223	0.4592
bruises	**0.6734**	0.1743	0.2532	0.1173
Odor	0.3324	0.4529	0.1284	0.4592
gill-attachment	0.4224	0.5243	**0.7239**	0.1934
gill-spacing	0.3479	0.6287	0.2737	0.2375
gill-size	0.4625	0.3634	0.2751	0.5467
gill-color	0.3624	0.7239	0.3587	0.6385
stalk-shape	**0.7859**	**0.6743**	0.2654	0.2765
stalk-root	**0.6739**	0.6498	0.2648	0.2389
stalk-surface-above-ring	0.3837	0.2374	0.2757	0.3850
stalk-surface-below-ring	**0.6824**	**0.6840**	0.3648	0.3742
stalk-color-above-ring	0.3747	0.5834	0.4577	0.3472
stalk-color-below-ring	0.4648	**0.7845**	0.4578	0.4855
veil-type	0.5673	**0.7393**	0.3476	**0.6734**
veil-color	0.3475	**0.7834**	0.3723	**0.7683**
ring-number	0.4738	0.3764	0.3463	0.6462
ring-type	0.3474	**0.7683**	0.4636	0.3753
spore-print-color	**0.6754**	**0.7463**	**0.6725**	0.2321
population	0.3401	**0.7385**	0.2647	0.2751
Habitat	0.3246	0.2346	0.2235	0.5204

Table 5.Clustering Accuracy of Proposed method with existing algorithms

Data size	CLIQUE(%)	DENCOS (%)	Enhanced DENCOS method(%)
1000	15	31	67
2000	28	40	64
3000	42	54	72
4000	48	60	76
5000	59	62	79
6000	61	75	80
7000	78	81	84

Table 5 shows the clustering accuracy for the subspace clustering algorithms CLIQUE, DENCOS and proposed enhanced DENCOS method. For the variation of data size between 1000 and 7000, it shows that the proposed method gives better accuracy when compared to existing methods.

Table 6.Performancecomparison with varied dimensions

Algorithm	Dimensions			
	25	50	100	150
CLIQUE	51	56	73	77
DENCOS	59	62	71	82
Proposed enhanced DENCOS	74	75	79	86

Table 6 gives the performance comparison for varied dimensions from 25 to 150. The results show that proposed method presents reliable results when compared to existing methods.

Table 7. Comparison of Precision, Recall and F-measure for Subspace clustering algorithms on breast cancer, heart disease and mushroom dataset

Algorithm	Breast cancer			Heart disease			Mushroom		
	Precision	Recall	F-Measure	Precision	Recall	F-Measure	Precision	Recall	F-Measure
CLIQUE	0.6316	0.6109	0.8126	0.5698	0.7343	07739	0.7429	0.7293	0.7494
DENCOS	0.7832	0.6281	0.8739	0.6901	0.8496	0.7948	0.7682	0.7566	0.7938
Proposed DENCOS	0.9101	0.8288	0.9213	0.8892	0.9029	0.8682	0.8605	0.8592	0.9183

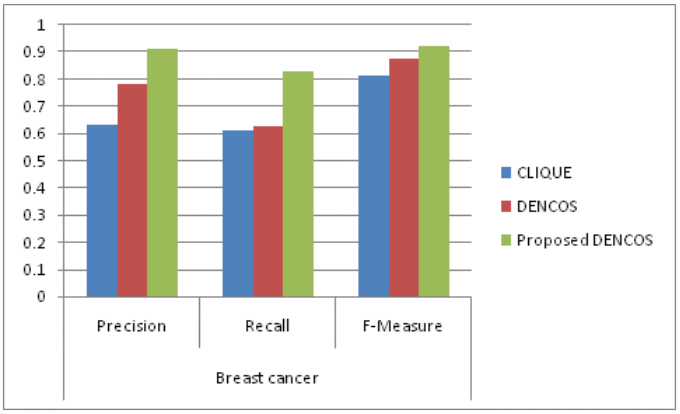

Figure 2. Graph of Precision, Recall and F-measure for CLIQUE, DENCOS, proposed DENCOS method for breast data set

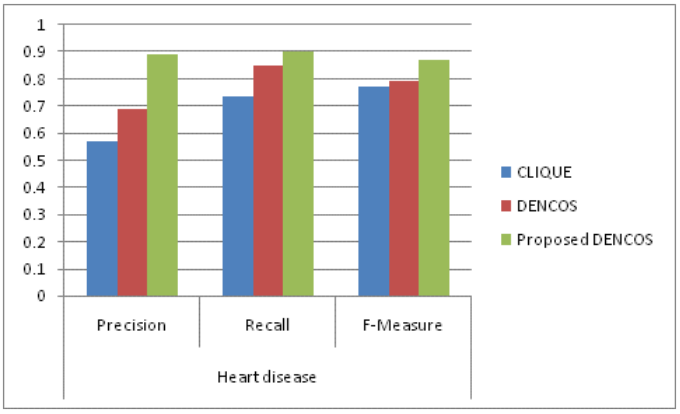

Figure 3. Graph of Precision, Recall and F-measure for CLIQUE, DENCOS, proposed DENCOS method for heartdisease data set

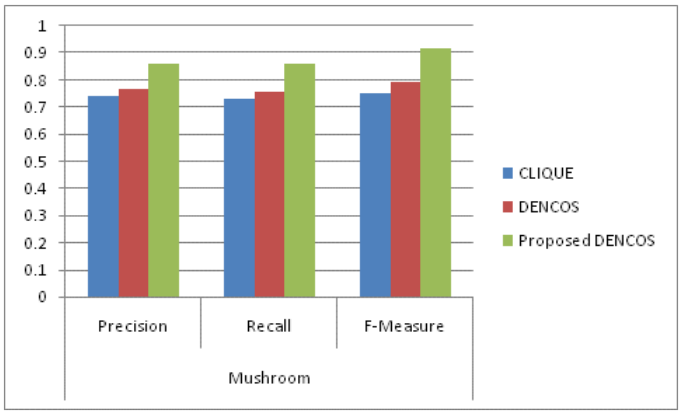

Figure 4. Graph of Precision, Recall and F-measure for CLIQUE, DENCOS, proposed enhanced DENCOS method for mushroom data set

Table 5. CLIQUE, DENCOS and proposed DENCOS algorithms with subspace dimensions for breast data set

Cluster No.	CLIQUE	DENCOS	Proposed DENCOS
1	1,2,3,4,5,6,7,8,9	2,5,7	2,7
2	1,2,3,4,5,6,7,8,9	3,4,7	3,4,7
3	1,2,3,4,5,6,7,8,9	5,7,8	5,7
4	1,2,3,4,5,6,7,8,9	2,4,5	2,4
5	1,2,3,4,5,6,7,8,9	3,7,9	5,7

6. Conclusion

In this paper an enhanced DENCOS approach is presented that identifies the relevant subspaces based on the attributes identified using fuzzy clustering method. FP-tree construction paves for identifying dense units of data. The fuzzy attribute weight has identified the attributes which are highly relevant to form subspace. For example, in breast cancer data set the attributes clump thickness, bare nuclei, normal nucleolus have a high probability that belong to cluster 1, which determines the severity in the formation of cancer cells. The accuracy of the proposed method is better when compared to CLIQUE and DENCOS method while the data size is 6000 and 7000. When the number of dimensions is 25 the performance of the proposed DENCOS method is considerably better when compared to CLIQUE method. The performance measures in terms of precision, recall and F-measure for the proposed model is significantly better when compared with the CLIQUE method.

References

1. J. C. Bezdek, Pattern recognition with fuzzy objective function algorithms, Plenum Press, New York, 1981.
2. K. Jain, Data clustering: 50 years beyond K-means, Pattern Recognition Letters, 31 (2010),651-666
3. L. Parsons, E. Haque, H. Liu, Subspace clustering for high dimensional data: A review, ACM SIGKDD Explorations Newsletter, 6 (2004), 90-105.
4. H. P. Kriegel, P. Kroger, A. Zimek, Clustering high-dimensional data: A survey on subspace clustering,

pattern-based clustering, and correlation clustering, ACM Transactions on Knowledge Discovery from Data, 3 (2009), 1-58.

5. E. Muller, S. Gunnemann, I. Assent, T. Seidl, Evaluating clustering in subspace projections of high dimensional data, in: Proc. VLDB Endowment, 2009, pp. 1270-1281.
6. Z. H. Deng, K. S. Choi, F. L. Chung, S. Wang, Enhanced soft subspace clustering integrating within-cluster and between-cluster information, Pattern Recognition, 43 (2010), 767-781.
7. L. Jing, M. K. Ng, J. Z. Huang, An entropy weighting k-means algorithm for subspace clustering of high-dimensional sparse data, IEEE Transactions on Knowledge and Data Engineering, 19 (2007), 1026-1041.
8. J.C. Dunn: A fuzzy relative of the ISODATA process and its use in detecting compact, well separated clusters, Journal Cybernetics, 3(3):32–57, 1973.
9. Krzysztof Simi´Nski, 2012, Clustering In Fuzzy Subspaces, Theoretical And Applied Informatics, Vol.24 (2012), No. 4, Pp. 313–326.
10. Lin Zhu, Jingsheng Lei, Zhongqin Bi, FeifeiXustudy Of Fuzzy Weighting Subspace Clustering For Streaming Data Journal Of Computational Information Systems 10: 14 (2014) 6305–6314
11. MdAnisurRahman And MdZahidul Islam, Crudaw: A Novel Fuzzy Technique For Clustering Records Following User Defined Attribute Weights, Proceedings Of The Tenth Australasian Data Mining Conference (Ausdm 2012) pp. 27-41.
12. Hongyan Liu And Feng Kong, 2005,'A New Madm Algorithm Based On Fuzzy Subjective And Objective Integrated Weights',International Journal Of Information And Systems Sciences, Volume 1, Number 3-4, Pages 420–427
13. Bezdek, J.C., 1974a. Numerical taxonomy with fuzzy sets. J. Math. Biology 1, 57–71
14. Bezdek, J.C., 1974b. Cluster validity with fuzzy sets. J.Cybernit. 3, 58–72
15. J.C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum Press, NewYork, 1981
16. A.M. Bensaid, L.O. Hall, J.C. Bezdek, L.P. Clarke, M.L. Silbiger, J.A.Arrington, and R.F. Murtagh. Validity-guided (Re)Clustering with applications to image segmentation,*IEEE Transactions on Fuzzy Systems*, 4:112{123, 1996.
17. XB-X. L. Xie and G. A. Beni. Validity measure for fuzzy clustering. IEEE Trans. PAMI, 3(8):841-846, 1991.
18. Dunn, J. C. "Well separated clusters and optimal fuzzy partitions", J. Cybern. Vol.4, pp. 95-104, 1974.