The Application of Improved GG Clustering Algorithm in View-irrelevant Behavior Recognition

Yun Liu, Jin Shao, Yan Yue

Qingdao University of Science and Technology, China

Keywords: GG clustering algorithm, Number of clusters, Cluster validity index, Behavior recognition.

Abstract. When cluster descriptors of behavior feature in the analyzing the behavior feature data of behavior under different view, the traditional FCM algorithm can not determine the number of clusters to the data with spherical structure, so this paper proposes an improved GG clustering algorithm to solve this problem. This algorithm determine the optimal cluster number by the indexes of inter-cluster compactness and the separation of clusters. Then model behavioral descriptors that have been clustered to reach the purpose of improving behavior recognition accuracy. The experimental results show that: the improved algorithm can classify and model behavioral descriptors better and improve the recognition accuracy.

Introduction

There are many different methods about behavior recognition and understanding in recent years[1-3],in the method of extracting behavior features, establishing behavior feature descriptor and analyzing the descriptors with clustering method to build behavior recognition model to complete the behavior recognitionthe effect of clustering and modeling has an important influence on behavior recognition.

Over the years many researchers has study deeply with the theoretical basis of validity index of clustering, study and improve the basic requirements the validity index of clustering should satisfy, and proposed a set of basic axiom that the validity index of clustering must satisfy. In 1965, the founder of fuzzy set theory Zadeh proposed a validity function of clustering: Separation function, but the judgment of fuzzy clustering validity is not very ideal. In 1974, Bezdek put forward the concept of partition coefficient, it constitutes the first practical clustering validity index PC[4], and then the concept of partition entropy is proposed PE[5]. In 1987, Davies and Bouldin[6] proposed separability measure based on Fisher distance between one cluster and another. In1991, Xie et al.[7] used objective function of fuzzy clustering, along with two important factors separation and compactness, proposed Xie-Beni index, but the evaluation standard does not consider the structure of data set. Kim et al.[8] proposed validity Kim based on overlapping degree among clusters, but the same as partition coefficient and partition entropy, it will monotonously change with the increase of the cluster number. In 1998, Rezaee proposed using linear combination to zoom compactness and separation by scale factor, thus make up for the defects of the difference measurement to a certain degree[9]. Although the index has a great improvement on overall performance, the structure is very complex, and always gives out the result that opposite the facts.

Considering that XB index does not consider the structure of data set and Kim index monotonously changes with the increase of the cluster number, this article uses the sum of weighted square errors within the cluster to measure the compactness, and uses the couples of fuzzy clustering to measure the separation of clusters. So defines the validity index(CS) based on these two metrics, effectively overcoming the shortcoming of the traditional FCM clustering algorithms on determining the initial parameter. The experimental results show that the improved algorithm achieves a more stable and effective clustering result.

Traditional GG Clustering Algorithm

Gath-Geva(GG) algorithm is an improvement of FCM algorithm. Fuzzy C- means clustering algorithm can only reflect the standard distance norms of super spherical data structure, so the FCM algorithm is only suitable for data structure with the same shape and direction. Thus GG clustering algorithm uses distance measure based on fuzzy maximum likelihood estimation, and can detect and adapt to sample data with different shape, size, density, at the same time, it makes the clustering no longer be limited by the sample data distribution volume and can improve the accuracy of clustering.

Suppose $X=\{x_1,x_2,\cdots,x_n\}$ is a data set with n metadata, x_i is sample data with p dimension, fuzzy clustering divide the sample set X into c clusters (suppose $c_2,\cdots,c_j,\cdots,c_c$) according to fuzzy partition matrix $U=\left[u_{ij}\right]$, $u_{ij}\in[0,1]$ represents the degree that the sample x_i belonging to class j, and the sum of membership degree of sample x_i belonging to all classes is 1, according to minimum distance quadratic sum that the sample point to the cluster center, define the objective function:

$$J=\sum_{j=1}^{c}\sum_{i=1}^{N}(u_{ij})^m\left|x_i-c_j\right|^2/2$$

GG clustering algorithm uses distance measure based on fuzzy maximum likelihood estimation:

$$D(x_i,\mathrm{c}_j)=\frac{(\det(A_i))^{1/2}}{p_i}\exp(\frac{(x_i-c_j^{(t)})^T A_i^{-1}(x_i-c_j^{(t)})}{2}),\quad 1\le i\le N,1\le j\le c$$

A_i is the covariance matrix of class i, p_i is the selected prior probability of class i.

$$A_i=\frac{\sum_{k=1}^{n}(u_{ik})^m(x_i-c_j^{(t)})(x_i-c_j^{(t)})^T}{\sum_{k=1}^{n}(u_{ij})^m},\quad p_i=\frac{1}{n}\sum_{k=1}^{n}u_{ik}$$

The clustering center is:

$$c_j^{(t)}=\sum_{i=1}^{N}(u_{ij}^{(t-1)})^m x_i/\sum_{i=1}^{N}(u_{ij}^{(t-1)})^m\quad,\quad 1\le j\le c$$

Membership renewal function is:

$$u_{ij}^{(t)}=\frac{1}{\sum_{k=1}^{c}(D(x_i,\mathrm{c}_j)/D(x_i,\mathrm{c}_k))^{2/(m-1)}},\quad 1\le i\le N,1\le j\le c$$

Improved GG Clustering Algorithm

This algorithm is based on the original algorithm, it uses index to determine the best number of clusters, achieving automatic clustering of traditional GG clustering algorithm.

Validity index (CS). Definition 1: The separation of clusters. In the fuzzy clustering division, in order to get the better clustering result, we should make the overlapping sample data fewer and the separation of clusters higher, so the separation of clusters can also be represented as the overlap of clusters.

$X=\{x_1,x_2,x_3,\ldots,x_n\}$ is the data set with n metadata. S_{c1} and S_{c2} are two fuzzy clusters belong to fuzzy division (U,V). The membership of sample x_i in S_{c1} and S_{c2} are represented

as $S_{c1}(x_i)$ and $S_{c2}(x_i)$ respectively. The total number that the clusters overlapping is N, and $N = c(c-1)/2$, c is the number of clusters.

Hence, the overlap of sample x_i in S_{c1} and S_{c2} is:

$$S(S_{c1}, S_{c2} : x_i) = \min(S_{c1}(x_i), S_{c2}(x_i))$$

The overlap of Fuzzy clusters S_{c1} and S_{c2} is:

$$S(S_{c1}, S_{c2}) = \sum_{k=1}^{N} S(S_{c1}, S_{c2} : x_k) \times \omega(x_k)$$

And in this formula $\omega(x_i) = -\sum_{i=1}^{c} u_{s_k}(x_k) \log_a u_{s_k}(x_k)$ .The weight index is used to adjust the overlapping part of data points, to weaken the effect of overlapping of a cluster when clustering. Therefore, the separation of clusters can be defined as average overlap. Because each two clusters may exists overlap, so the average overlap can be expressed as the formula:

$$S_{重}(c,U) = \frac{2\sum_{i \neq j}^{c} S(S_{ci}, S_{cj})}{c(c-1) \times n} \quad , S(c,U) = 1 - S_{重}(c,U)$$

From the formula we can see the lower the overlap is in clustering, the higher the separation of clusters is. So, the higher the separation of clusters is, namely the higher $S(c,U)$ is, the better the clustering result is, and get the optimal cluster number.

Definition 2： Inter-cluster compactness. In fuzzy division, the compactness of data is represented as the sum of weighted square errors within the cluster, it is defined as follows:

$$C(c,U) = \sum_{i=1}^{c} \frac{\sum_{j=1}^{N} (u_{ij})^m \left\| x_j - v_i \right\|^2}{n_i} * (\frac{c+1}{c-1})^{1/2}$$

Where $\sum_{j=1}^{N} (u_{ij})^m \left\| x_j - v_i \right\|^2$ is the sum of square errors within the cluster based on the Euclidean distance; Fuzzy cardinality $n_i = \sum_{j=1}^{N} u_{ij}$, decreases with the increase of c, so as the weight of each cluster, $\frac{1}{n_i}$ limits the monotone decreasing of compactness measure.When the value of $Com(c,U)$ reaches the minimum, the compactness is best, and the inter-cluster aggregation reaches the best.

Definition 3：Validity index (CS). Because the separation and compactness have different measure standard, so it needs normalization processing. The results can be expressed as:

$$Com(c,U) = \frac{C(c,U)}{\max\{C(c_1,U),...,C(c_{\max},U)\}}, \quad Sep(c,U) = \frac{S(c,U)}{\max\{S(c_1,U),...,S(c_{\max},U)\}}$$

Validity index CS can be expressed as: $CS = \frac{Com(c,U)}{Sep(c,U)}$

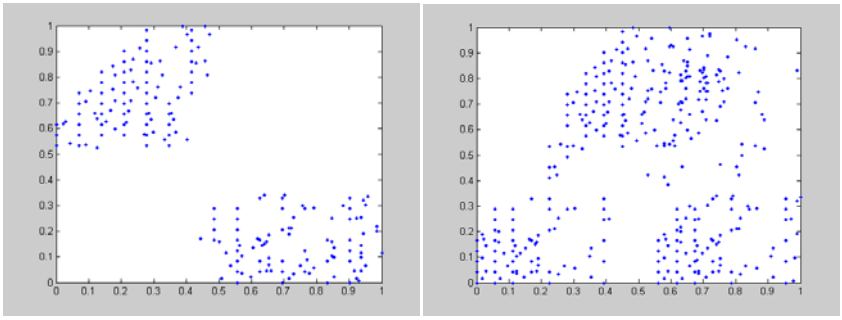
In summary, when the separation of clusters is lower and the inter-cluster compactness is higher, the value of CS is also lower. Therefore, the number of clusters is best when the value of CS reaches the minimum.

Experimental Results and Analysis

To verify the effectiveness of GG clustering algorithm, we use the data set to test the

effectiveness of index CS, according to the past experience, fuzzy index m generally is a number between 1.5 and 2.5, the algorithm supposed m = 2.

The experimental data sets. The experiment used two groups of artificial data sets and one group of real data set (IRIS).



Data Set 1 Data Set 2

Fig. 1: Artificial Data Sets

Clustering experimental analysis of validity index.In order to verify the validity of index CS, compared it with four common indexes. The four common indexes are partition coefficient PC, partition entropy PE, classic validity index XB, relative index DI. The experiment processed three data sets using GG clustering algorithm. Table 1, Table 2 and Table 3 list the values of each index respectively of data set 1, data set 2 and data set 3(IRIS data set) when the number of clusters changes from 2 to 8. With the difference of clustering number, the values of each index are also different, When PC reaches the maximum, PE reaches the minimum, XB reaches the minimum, DI reaches the maximum, the number of clusters reaches the best, and CS reaches the minimum the number of clusters reaches the best.

Table 1. Result of artificial data set 1

Clustering number	M(PC)	M(PE)	M(XB)	M(DI)	M(CS)
2	0.9127	0.1735	1.6593	0.4315	0.7435
3	0.8015	0.3890	1.7131	0.0739	0.8224
4	0.7591	0.4548	2.2557	0.0626	0.9167
5	0.6338	0.6329	3.9628	0.0376	0.9023
6	0.6125	0.7297	1.7222	0.0431	0.9801
7	0.5956	0.8001	3.4577	0.0573	0.9890
8	0.5726	0.8975	5.5482	0.0617	0.9918

Table 2. Result of artificial data set 2

Clustering number	M(PC)	M(PE)	M(XB)	M(DI)	M(CS)
2	0.8166	0.4011	1.7547	0.0688	0.8802
3	0.7913	0.4579	1.6908	0.0381	0.7911
4	0.7254	0.5228	2.8534	0.0292	0.8287
5	0.6517	0.6946	1.7116	0.0257	0.8592
6	0.5932	0.8127	1.8109	0.0154	0.9428
7	0.5413	0.8302	4.3372	0.0178	0.9720
8	0.4902	0.9567	3.0604	0.0239	1.0949

Table 3. Result of IRIS data set 3

Clustering number	M(PC)	M(PE)	M(XB)	M(DI)	M(CS)
2	0.8316	0.2134	3.4943	0.3581	0.7370
3	0.7833	0.3769	2.7263	0.0938	0.7209
4	0.6257	0.5951	4.4212	0.0560	0.7636
5	0.5902	0.6683	3.0349	0.0761	0.8197
6	0.5681	0.8224	2.9936	0.0444	0.8445
7	0.5524	0.9412	3.4079	0.0701	0.9023
8	0.4463	0.9846	5.6643	0.0586	0.9576

Table 1, Table 2 and Table 3 show the experimental result: Partition coefficient PC and PE only predict the right number of clusters on the data set 1 whose data are definitely separated. Index

XB predicts the right number of clusters on the data set 1 and data set IRIS. Index DI predicts the right number of clusters on data set 1. Index CS predicts the right number of clusters on the three data sets.

In summary of the experiment, CS index can not only predict the accurate number of clusters, but also can properly divide the data set that there is overlap between the classes, and it can measure data sets with multiple geometries.

Behavior Recognition Experiment

This experiment applied GG clustering algorithm combined of this index to the behavior recognition and clustered the behavioral recurrence descriptors that extracted from behavior in video, and then matched the behavioral descriptors of behavior for test to the classification models, to determine whether the behavior is same to the behavior that the template represents.

In this experiment, we used video data of IXMAS multi-view video database, and extracted recurrence characteristic descriptors of behavior kick, then got different classification models by changing the number of fuzzy clustering, and matching ten segments of videos to classification models that established with different indexes, to verify the validity of the index. And the threshold parameter of behavior recognition is 0.2, namely when the distance between test sample and template is less than 0.2, considered the behavior of test sample was considered the same kind of behavior as the behavior the template represents. Table 4 lists the values of CS with different number of different clusters. Table 5 lists the recognition validity of ten segments of videos matched to classification models, after modeling with different number of clusters.

It can be seen from Table 4, when the number of clusters is 8, the CS value reaches the optimum, meanwhile Table 5 shows the recognition rate is highest by using the model built with this number of clusters. The experimental result shows that this index has great advantage in the modeling of behavior recognition in videos, and provides guarantee for building model correctly

Tab.4 Different CS values under different cluster number

Clustering number	M(CS)
4	0.8523
5	0.6462
6	0.8219
7	1.1047
8	0.5371
9	0.7743
10	0.7954

Tab.5 Video identification number under different number of clusters

Clustering number	The number of test video	Video identification number	The recognition rate
4	10	5	50%
5	10	7	70%
6	10	5	50%
7	10	4	40%
8	10	7	70%
9	10	5	50%
10	10	6	60%

As can be seen from table, when the number of clusters is 8, the value of CS to achieve, at the

same time, in the video, the highest recognition rate. The experimental results show that, the index has great advantages in modeling video action recognition, establish the correct model provides a guarantee.

Conclusions

This paper applies the index based on separation of clusters and inter-cluster compactness to the GG clustering algorithm. It can be seen from the analysis of experimental data that the index can indicate the best number of clusters effectively. Last, apply this improved algorithm to the modeling for behavior recognition in videos. The experimental result shows that the improved algorithm can indicate the best number of clusters indicate the optimal number of clusters in clustering and modeling of behavior recognition, and improve the accuracy of View-irrelevant behavior recognition, so in terms of accuracy requirement, the index has good applicability.

Acknowledgements

The authors would like to thank the Internet of Things and Intelligent Information Laboratory which is the key laboratory of Qingdao University of Science and Technology. This research is also partially supported by National Fund of Abnormal Behavior Detection and Real-time Transmission Resaerch in Video Surveillance under Cloud Computing Model (61142003) and the Research of Multi-scale Feature Calculation Accelerating Algorithm and View-irrelative Descriptors Mining Method of Same Kind of Behavior (61472196).

References

[1] Chen Changhong, Liu Zhijing, "the crowd behavior analysis research, computer science, 39 (10): 7-112012

[2] Tian Lan,Leonid Signal,Greg Mori,“Social Roles in Hierarchical Models for Human Activity Recognition”,ECCV:4321-4328.,2012

[3] Pyry Matikainen,Rahul Sukthankar,Martial Hebert,“Model recommendation for action recognition”,CVPR 2012:2256-2263.

[4] Bezdek J C. Cluster Validity with Fuzzy Sets[J].Journal of Cybernetics, 1974,3(3):58-73.

[5] Bezdek J C. Numerical Taxonomy with Fuzzy Sets[J]. Math Biol,1974, 1(1): 57-71.

[6] D.L. Davies, D.W. Bouldin, A Cluster Separation Measure, IEEE Trans Syst Man Cyber, 1987, 17: 873-877

[7] XIE X L, Beni G, A validity measure for fuzzy clustering, IEEE Trans, on Pattern Analysis and Machine Intelligence, 1991, 13(8): 841~847

[8] KIM D W， LEE K H， LEE D． On cluster validity index for estimation，of the optimal number of fuzzy clusters ［J］． Pattern Recognition，2004， 37(10) : 2009-2025．

[9] Jrezaee M,Letlieveldt B,ReiberJ.A new cluster validity index for the FuZZy c-means.Pattern Recognition Letters, 1998，19:237-246