Extension to C-means Algorithm for the Use of Similarity Functions

Javier Raymundo García-Serrano[1] and José Francisco Martínez-Trinidad[2]

[1]Centro Nacional de Investigación y Desarrollo Tecnológico. Cuernavaca, Morelos, México.
e-mail: acjrgs@cenidet.edu.mx
[2]Centro de Investigación en Computación Instituto Politécnico Nacional, México, D.F.
e-mail: fmartine@pollux.cic.ipn.mx

Abstract. The C-Means algorithm has been motive of many extensions since the first publications. The extensions until now consider mainly the following aspects: the selection of initial seeds (centers); the determination of the optimal number of clusters and the use of different functionals for generate the clusters. In this paper it is proposed an extension to the C-means algorithm which considers description of the objects (data) with quantitative and qualitative features, besides consider missing data. These types of descriptions are very frequent in soft sciences as Medicine, Geology, Sociology, Marketing, etc. so the application scope for the proposed algorithm is very wide. The proposed algorithm use similarity functions that may be in function of partial similarity functions consequently allows comparing objects analyzing subdescriptions of the same. Results using standard public databases [2] are showed. In addition, a comparison with classical C-Means algorithm [7] is provided.

1 Introduction

The restricted unsupervised classification (RUC) problems have been studied intensely in the Statistical Pattern Recognition [1,6] based on metric distances in an n-dimensional metric space [1,6]. Algorithms as the C-means have showed their effectiveness in the unsupervised classification process. This algorithm starts with an initial partition, then tries all possible moving or swapping of data from a group to another iteratively to optimize the objective measurement function. The objects must be described in terms of features such that a metric can be applied for evaluate the distance. Nevertheless, these are not the conditions in soft sciences as Medicine, Geology, Sociology, Marketing, etc. [4]. In this sciences the objects are described in terms of quantitative and qualitative features. For example, if we look at geological data, features such as age, porosity, and permeability are quantitative, but other features such as rock types crystalline structure, and facies structure are qualitative. Likewise, the missing data is common in this type of problems. In this circumstances only the similarity degree between the objects can be determined. Actually exists few algorithms for solve the RUC in a context as the previously mentioned. The conceptual C-means algorithm is the most representative [3]. This algorithm proposes a distance function for handle quantitative and qualitative features. The distance between two objects is computed evaluating the distance between quantitative

J.M. Zytkow and J. Rauch (Eds.): PKDD'99, LNAI 1704, pp. 354-359, 1999.

features (with an Euclidean distance) plus the distance between qualitative features (using the chi-square distance). For achieves before, each value of a qualitative feature is coding as a binary feature. Finally is assumed that the distance defined in this way has an interpretation in the original n-dimensional space, where centroides for example can be computed, this last is wrong.

Other motivation for our proposed algorithm is the necessity of many specialists, that work in soft sciences, for group data in a specific number of clusters (a RUC problem). Such that objects that are more similar tend to fall into the same group and objects that are relatively distinct tend to separate into different groups.

2 Algorithm Description

Let us consider a set of m objects $\{O_1,O_2,...,O_m\}$ which must be grouped in c clusters. Each object is described by a set $R=\{x_1,x_2,...,x_n\}$ of features. The features take values in a set of admissible values $x_i(O_j)\in M_i$, $i=1,...,n$. We assume that in M_i exists a symbol „*“ for denote missing data. The features thus can be of any nature (qualitative: Boolean, multi-valued, etc. or quantitative: integer, real) and incomplete descriptions of the objects can be considered. For each feature a comparison criteria $C_i:M_i x M_i \rightarrow L_i$ $i=1,\ldots,n$ is defined, where L_i is a totally ordered set, besides, Let $\Gamma:(M_1\times...\times M_n)^2\rightarrow[0,1]$ be a similarity function. In some cases the similarity function Γ depend or is a lineal combination of partial similarity functions $\Gamma':(M_{i_1}\times...\times M_{i_s})^2\rightarrow L_i'$, with L_i' same as L_i. This function allow us to compare descriptions of objects with $s<n$. A subset non empty of features for analyze the sub-descriptions of objects is denominated *support set*. A set formed by support sets is denominated *support sets system*. From now on, we will use O_i instead of $I(O_i)=(x_1(O_j),\ldots,x_n(O_j))$ for denote the description of an object. Let $\Gamma(O_j,O_k)$ the similarity between the objects O_j and O_k. The value $\Gamma(O_j,O_k)$ satisfies the following three conditions:

1. $\Gamma(O_j,O_k)\in[0,1]$ for $1\le j\le m$ and $1\le k\le m$;
2. $\Gamma(O_j,O_j)=1$ for $1\le j\le m$;
3. $\Gamma(O_j,O_k)=\Gamma(O_k,O_j)$ for $1\le j\le m$ and $1\le k\le m$.

Let u_{ik} the degree of membership of the object O_k in the cluster C_i, and let R^{cxm} the set of all the real $c\times n$ matrices. Any c-partition of the data set is represented by a matrix $U=[u_{ik}]\in R^{cxm}$, which satisfies:

1. $u_{ik}\in\{0,1\}$ for $1\le j\le m$ and $1\le k\le m$;
2. $\sum_{i=1}^{c} u_{ik}=1$ for $1\le k\le m$;
3. $\sum_{k=1}^{m} u_{ik}>0$ for $1\le i\le c$.

Then we have that, the partition matrix U is determined from maximization of the objective function given by $J(U)=\sum_{i=1}^{c}\sum_{k=1}^{m} u_{ik}\Gamma(O_i^r,O_k)$ where $\Gamma(O_i^r,O_k)$ is the

similarity between the object most representative O_i^r („the center“) in the cluster C_i and the object O_k. Note that in our case the „center“ is an object of the sample instead of a fictitious element as in the classical C-means algorithm. To determine the most representative object in a cluster C_i for a given U, we introduce the function

$$r_{\substack{C_i \\ O_j \in C_i \\ C_q \neq C_i}}(O_j) = (\beta_{C_i}(O_j) / [(\alpha_{C_i}(O_j) + (1 - \beta_{C_i}(O_j)))]) + \eta_{C_q}(O_j) \quad (1)$$

where the function $\beta_{C_i}(O_j)$ is the average of similarity (mean) of the object O_j with the other objects in the same cluster C_i and is computed as follows

$$\beta_{C_i}(O_j) = \frac{1}{|C_i| - 1} \sum_{\substack{O_j, O_q \in C_i \\ O_j \neq O_q}} \Gamma(O_j, O_q) \quad (2)$$

For increase the informational value of (2) we introduce the function $\alpha_{C_i}(O_j)$.

$$\alpha_{C_i}(O_j) = \frac{1}{|C_i| - 1} \sum_{\substack{O_j, O_q \in C_i \\ O_j \neq O_q}} |\beta_{C_i}(O_j) - \Gamma(O_j, O_q)| \quad (3)$$

This function evaluates the variance between the mean (2) and the similarity between the object O_j and the other objects in C_i, so when the variance decreases increases the values of (1). To the expression (3) in (1) is added the expression $(1 - \beta_{C_i}(O_j))$ in the denominator. This value allow us compute (1) in the case that (3) be zero and besides represents the average of dissimilarity of O_j with the other objects in C_i.

$$\eta_{C_k}(O_j) = \sum_{\substack{q=1 \\ i \neq q}}^{c} (1 - \Gamma(O_q^r, O_j)) \quad (4)$$

Finally the function (4) is used for diminish the cases where exist two objects with the same value in (1). The function (4) represents the dissimilarity of the object O_i with the representative objects of the others clusters.

It is quite reasonable that the representative object for the cluster C_i is defined as the object O_r which yield the maximum of the function $r_{C_i}(O_j)$.

$$r_{C_i}(O_r) = \max_{O_p \in C_i} \{r_{C_i}(O_p)\} \quad (5)$$

Therefore, the most representative object is such that, it is the most similar in average with the other objects in the cluster, being this average the greater and with less variance. In addition, the representative object is the most dissimilar compared with the representative objects in the other clusters.

If the cluster centers are given, the functional $J(U)$ is maximized when u_{ik} is determined as:

$$u_{ik} = \begin{cases} 1 & \text{if } \Gamma(O_i^r, O_k) = \max_{1 \leq q \leq c} \{\Gamma(O_q^r, O_k)\} \\ 0 & \end{cases} \quad (6)$$

That is to say, an object O_k will be assigned to the cluster such that O_k is the more similar with their representative object.

2.1 Algorithm

Step 1. Select *c* objects in the data as initial seeds. Fix the number of iterations *ni*', and *ni*=0.
Step 2. Calculate the partition matrix $U=U^{(ni)}$ using (6) and the initial seeds selected in the step 1.
Step 3. Determine the representative objects of the clusters for the matrix $U^{(ni)}$, using (1) and (5).
Step 4. Calculate the partition matrix $U^{(ni+1)}$ using (6) and the representative objects computed in the step 3.
Step 5. If the set of representative objects, is the same that in the previous iteration stop. Otherwise increase *ni=ni+1*.
Step 6. If *ni>ni'* stop. Otherwise, go to step 3.

3. Experimental Results

Initially we perform a comparison between our extension and the classical c-means algorithm, for this purpose we consider the Iris data [2]. We apply the classical c-means algorithm using the Euclidean distance, and our extension using the equations (7) as comparison criteria between feature's values, and equation (8) as similarity function between object's descriptions. The results are showed in the table 1.

$$C_s\left(x_s(O_i), x_s(O_j)\right) = 1 - \left|1 - \left(1/\left(\left|x_s(O_i) - x_s(O_j)\right| + 1\right)\right)\right| \quad (7)$$

$$s(O_i, O_j) = 1/|T| \sum_{x_p \in T} C_p\left(x_p(O_i), x_p(O_j)\right) \quad (8)$$

Table 1. Results using classical C-means and the proposed extension in 10 realized tests

Type of initial seeds	% Effectiveness per cluster			Used Algorithm	% Effectiveness in average
	Cluster1	Cluster2	Cluster3		
Random	62%	8%	100%	Classic C-means	56.7 %
Random	99%	75%	93%	Extension proposed	89.16 %
Representative[1]	100%	70%	96%	Classic C-means	88.6 %
Representative	99%	78%	92%	Extension proposed	89.8 %

So our extension obtains a better percent of classification in average than the classical c-means algorithm, this percent could be improved by a better modeling of (7) and (8).

[1] An initial representative seed is an object selected from the cluster to form.

Finally, we test the algorithm selecting three databases from [2]. In these databases is known their arrangement in order to evaluate the percentage of correct classification. For both Iris and Win databases the comparison criteria used to compare the features' values was (7) and the similarity function used to compare the objects was (8). In the case of Mushroom Database was used the comparison criterion (9) for all the features and the similarity function used in this case was just (7).

$$C\left(x_s(O_i), x_s(O_j)\right) = \begin{cases} 1 & \text{if } x_s(O_i) = x_s(O_j) \\ 0 & \text{otherwise} \end{cases} \quad (9)$$

In the table 2 are presented the obtained results of apply the extension to C-means algorithm using random initial seeds.

Table 2. Results testing the extended C-means algorithm

Database	# Objects	# Quantitative features	# Qualitative features	# Clusters	# Tests	Missing Values	% Effectiveness
Iris	150	4	0	3	10	0	89.48%
Wine	178	8	5	3	10	0	86.3%
Mushroom	8124	22	0	2	10	2480	89.2%

In the Table 3 are showed the main differences between the classical C-means algorithm and our proposal.

Table 3. Differences between the classical algorithm c-means and the proposed extensions

Classical C-means	*Extensions proposed to C-Means*
It is metric	It is not metric
It is based on the Euclidean distance	Use comparison criteria and function of similarity
It works only with quantitative numerical descriptions	It works with mixed descriptions (quantitative and qualitative features)
It does not consider missing data	It considers missing data
It does not consider comparing subdescriptions	It considers comparing subdescriptions in base to a support set.

4 Conclusions

In this work, the C-means algorithm for the use of similarity functions in crisp case is proposed to solve RUC problems. The algorithm considers descriptions of the objects with mixed data, i.e. quantitative and qualitative features. Besides missing data is supported by the algorithm. This characteristics allows to the algorithm be potentially useful in many problems of Data Mining and knowledge Discovery.

In comparison with the classical C-means algorithm, our proposal presents in average a better classification than classical C-means using the Iris data. Besides, it allows analyze objects described with qualitative and quantitative features and missing data. Therefore, the proposed algorithm can be applied in fields as Medicine, Marketing, Geology, and Sociology, in general in the named soft sciences where the specialists face with this type of descriptions.

The use of comparison criteria by feature and their integration in a similarity function allows modeling more precisely a problem. In this way the expert's knowledge in soft sciences can be put in computer systems for solve data analysis and classification problems.

If the similarity function is a lineal combination of partial similarity functions, and support sets are used. Then the function allows a better discrimination between the clusters because the comparison between the objects is realized considering subsets of features emphasizing the similarities and differences between the objects allowing to the algorithm determining a better solution.

5 Future Work

The C-means algorithm is an iterative algorithm, which base their operation in the initial seeds so as future work we will propose a method for select candidate to initial seeds.

We are developing an optimal algorithm that can be applied for solve problems with big bulk of data.

Finally, an extension of our algorithm in the fuzzy case will be proposed in the future.

Acknowledge.-This work was partially financed by Dirección de Estudios de Posgrado e Investigación IPN and the CONACyT Project REDII99 Mexico.

References

1. Richard O. Duda and Peter E. Hart. Pattern Classification and Scene Analysis. (USA, John Wiley & Sons, Inc. 1973.).
2. ftp://ftp.ics.uci.edu/pub/machine-learning-databases/.
3. H. Ralambondrainy. „A conceptual version of the K-means algorithm". In: Pattern Recognition Letters 16, p. 1147-1157.
4. José Ruiz Shulcloper, et al. Introducción al reconocimiento de patrones (Enfoque Lógico Combinatorio) (Serie Verde No. 51, México, Depto. de Ingeniería Eléctrica, Sec. Computación CINVESTAV-IPN, 1995).
5. Ruspini, E. R. „A new approach to clustering" (En: Information and control, No. 15, 1969) p 22-32.
6. Robert J. Schalkoff. Pattern Recognition: Statistical, Structural and Neuronal Approaches (USA, John Wiley & Sons, Inc. 1992).
7. G. Ball and D. Hall, A Clustering technique for summarizing multivariate data, Behav. Sci., vol. 12, pp. 153-155, 1967.