http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet

Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 6 (2): 251-254, 2007
ISSN 1812-5638

Study on Mutual Information Based Clustering Algorithm

[1,2]Hongfang Zhou, [1]Boqin Feng, [2]Lintao Lv and [1]Hui Yue
[1]School of Electronics and Information Engineering, Xi'an Jiaotong University,
Xi'an, 710049, China
[2]School of Computer Science and Engineering, Xi'an University of Technology,
Xi'an, 710048, China

Abstract: Traditional clustering algorithms are designed for isolated datasets. But in most cases, relationships among different datasets are always existed. So we must consider the actual circumstances from the cooperative aspects. A new collaborative model is proposed and based on this model a new cooperative clustering algorithm is presented. In theorem, the algorithm is proved to converge to the local minimum. Finally, experimental results demonstrate that the clustering structures obtained by new algorithm are different from those of conventional algorithms for the consideration of collaboration and the performances of these collaborative clustering algorithms can be much better than those traditional separated algorithms under the cooperating circumstances.

Key words: Mutual information, clustering, cooperative model, fuzzy algorithm, global analysis, local analysis

INTRODUCTION

Clustering can be used in many different research domains (Han *et al*., 2001). But it is not suitable on any circumstance. They can either only cope with the isolated or independent datasets, or need surprising run time (Kanungo *et al*., 2002; Topchy *et al*., 2002). Especially with the web information crazy increasing, it is evident to be disabled. In view of these situations, the new clustering algorithms are urgent to be built. On the analysis large volume of conventional research data, we present a mutual information based approach to measure the cooperative relations among several datasets. And we discuss a new cooperative model. Furthermore, based on this we introduce corresponding cooperative clustering algorithm MICCA. Experiment results show that it is feasible and effective.

Fuzzy clustering algorithm: The most famous fuzzy clustering algorithm is FCM (Fuzzy C-Means) (Hopper, 1999) and its objective function JFCM is given by:

$$J(B,U;X) = \sum_{k=1}^{K}\sum_{j=1}^{C}\mu_{kj}^{m}d_{kj}^{2} \qquad (1)$$

In this formula, K is the number of samples; C is the number of clusters; m∈ [1,∞) is a weighting exponent called fuzzier d_{kj}^2 and is the distance from feature point x_k to prototype β_j. Minimization of (1) is usually achieved by a Picard iteration technique, which updates memberships and prototypes in an alternating fashion until convergence. The K×C matrix $U = [\mu_{kj}]$ is called a constrained fuzzy C-partition matrix if it satisfies

$$\mu_{kj} \in [0,1] \text{ for all i, j, } 0 < \sum_{k=1}^{K}\mu_{ij} < N \text{ for all i,} \qquad (2)$$
$$\text{and } \sum_{i=1}^{C}\mu_{ij} = 1 \text{ for all j}$$

In general, d^2 (x_k, v_j) corresponds to Euclidean measures. According to these principles, we can infer that FCM's membership update equation for this formulation is:

$$\mu_{kj} = \frac{\left(\frac{1}{\|x_k - v_j\|}\right)^{\frac{1}{m-1}}}{\sum_{i=1}^{C}\left(\frac{1}{\|x_k - v_i\|}\right)^{\frac{1}{m-1}}} \qquad (3)$$

And the center update equation is:

Corresponding Author: Hongfang Zhou, School of Electronics and Information Engineering, Xi'an Jiaotong University, Xi'an, 710049, China

$$v_j = \frac{\sum_{k=1}^{K} \mu_{kj}^2 x_k}{\sum_{k=1}^{K} \mu_{kj}^2} \tag{4}$$

Probability clustering algorithm: Given a dataset $\Gamma = [x_1, x_2,..., x_k\}$ and its corresponding clusters' centroids is v_j ($j = 1, 2,...,C$), then probability clustering objective function J is defined as follows:

$$\begin{aligned} J = J(\Gamma, V) &= \sum_{k=1}^{K} p(x_k) \sum_{j=1}^{C} p(v_j \mid x_k) \left\| x_k - v_j \right\|^2 \\ &= \sum_{k=1}^{K} \sum_{j=1}^{C} p(x_k) p(v_j \mid x_k) \left\| x_k - v_j \right\|^2 \\ &= \sum_{k=1}^{K} \sum_{j=1}^{C} p(x_k) \frac{\exp \sum_{i \in mfs(x_k, v_j)} \lambda_i f_i(x_k, v_j)}{\sum_{j=1}^{C} \exp \sum_{i \in mfs(x_k, v_j)} \lambda_i f_i(x_k, v_j)} \left\| x_k - v_j \right\|^2 \end{aligned} \tag{5}$$

In this formula,

$mfs\ (x_k, v_j) = \{i | f_i\ (x_k, v_j) \neq 0, 1 \le i \le n\}$.

Therefore, in order to compute $p\ (v_j \mid x_k)$, we need to find all of the possible cluster v_j containing x_k. And $mfs\ (x_k) = \cup mfs\ (x_k, v_j)$ corresponds the cluster collections containing x_k.

ALGORITHM

Mutual information: In this section, we will use a new information theory based mutual information measure to quantify cooperative relations among datasets (Shen *et al*., 2005). Considering simplicity and generality, we assume two datasets T and S. As described above, for every sample x_k, there exists a probability p_{kj} which denotes the possibility of x_k belonging to c_j.

Lemma 1: If only a measurement judging a sample's uncertainty information as to some specific cluster is used, then it is only relevant to p_{kj} expressing the probability of sample x_k belonging to cluster c_j. That is to say, there exists a function defined in the interval between 0 and 1, which satisfies

$$\begin{aligned} &(\text{"uncertainty informtion about sample } x_k \\ &\text{belong to cluster } c_j\text{"}) = f(p_{kj}) \end{aligned} \tag{6}$$

In Formula (6) p_{kj}^T andp_{kj}^S correspond to the probabilities of sample x_k belonging to cluster c_j in target dataset T and source dataset S, respectively. As described in Lemma 1, if dataset T is expressed by expected distribution

$\phi = \{ p_{kj}^S \mid k = 1,2,...,K; j = 1,2,...,C\}$, the function $f(p_{kj}^S\ )$

measures every sample's uncertainty in T constrained by expected distribution ϕ. Suppose

$p = \{ p_{kj}^T \mid k = 1,2,...,K; j = 1,2,...,C\}$

is the actual distribution of T, the Lemma 2 can be inferred.

Lemma 2: If the real probability $p_{1j}^T, p_{2j}^T, ..., p_{kj}^T$ distribution of T and discrete random variables

$f(p_{1j}^S), f(p_{2j}^S), ..., f(p_{kj}^S)$

are used to compute expectation $_p(\phi)$, $_p(\phi)$ is then corresponded to T's uncertainty information under the condition $\phi = \{\ p_{kj}^S\ \}$:

$$_P(\phi) = E_P[f(\phi)] \sum_{k=1}^{K} \sum_{i=1}^{C} p_{kj}^T f(p_{kj}^S) \tag{7}$$

Here, E_p is the mathematical expectation of P.
Def. 1. Suppose the real probability distribution of dataset T is

$P = \{p_{kj}^T \mid k = 1,2,...,K; j = 1,2,...,C\}$

and the real probability distribution of dataset S is

$\phi = \{p_{kj}^S \mid k = 1,2,...,K; j = 1,2,...,C\}$

S's mutual information gotten from T is:

$$\begin{aligned} \varepsilon(P, \phi) &= {}_P(\phi) - {}_P(P) \\ &= \sum_{k=1}^{K} \sum_{j=1}^{C} p_{kj}^T f(p_{kj}^S) - \sum_{k=1}^{K} \sum_{j=1}^{C} p_{kj}^T f(p_{kj}^T) \\ &= -\sum_{k=1}^{K} \sum_{j=1}^{C} p_{kj}^T [\ln p_{kj}^S - \ln p_{kj}^T] \end{aligned} \tag{8}$$

and we call $\varepsilon\ (P, \phi)$ as 's mutual information measurement gotten from T.

According to deduction above, we can generalize them and infer that if dataset S is affected by N datasets $n_1, n_2, ..., n_N$ at the same time. In this case, formula (8) can be transformed as:

$$\varepsilon(N,\phi) = -\sum_{\substack{i=1\\ l\neq S}}^{N}\sum_{k=1}^{K}\sum_{j=1}^{C} p_{kj}^{i}[\ln p_{kj}^{S} - \ln p_{kj}^{i}] \tag{9}$$

Cooperative clustering algorithm (MICCA) based on mutual information: In pervious sections, the dataset S's mutual information gotten from other datasets D_i in the cooperative environment is ε (N, ϕ) . Accordingly, under these circumstances, S's clustering objective function is:

$$\begin{aligned} J_C^S &= \sum_{k=1}^{K}\rho(\sum_{j=1}^{C} D_{jk}^{1/1-m})^{1-m} + \varepsilon(L,\phi) \\ &= \sum_{k=1}^{K}\rho(\sum_{j=1}^{C} D_{jk}^{1/1-m})^{1-m} - \sum_{\substack{i=1\\ i\neq S}}^{N}\sum_{k=1}^{K}\sum_{j=1}^{C} p_{kj}^{i}[\ln p_{kj}^{S} - \ln p_{kj}^{i}] \end{aligned} \tag{10}$$

since

$$(\sum_{j=1}^{C} D_{jk}^{1/1-m})^{1-m}$$

values corresponding to outliers are large, the idea is to design the objective function so that its global minimum is achieved when large

$$(\sum_{j=1}^{C} D_{jk}^{1/1-m})^{1-m}$$

are discounted or ignored. In formula (10), $\rho(\cdot)$ is a loss function to reduce the effect of outliers. The loss function is typically linear for small distances and then saturates for larger ones.

Because $\rho(\cdot)$ is a loss function and its function is reducing the effect of the noises. The loss function's value is increasing with respect to the small values until its independent variable get to some certain values. It is obvious that when $\rho(\cdot)$ reaches a constant its differential coefficient is close to 0. And it is reasonable to ignore it. That is to say, when $\rho(\cdot)$'s value is large enough, the clustering objective function J_C^S becomes

$$J_C^S = \varepsilon(N,\phi) = -\sum_{\substack{i=1\\ i\neq S}}^{N}\sum_{k=1}^{K}\sum_{j=1}^{C} p_{kj}^{i}[\ln p_{kj}^{S} - \ln p_{kj}^{i}] \tag{11}$$

When formula (11) is minimized, formula (12) is satisfied.

$$\frac{\partial J_C^S}{\partial v_j^S} = -2\sum_{k=1}^{K} p_{kj}^{S}(x_k^S - v_j^S) = 0 \tag{12}$$

Through formula (12), we can get:

$$v_j^S = \frac{\sum_{k=1}^{K} p_{kj}^{S} x_k^S}{\sum_{k=1}^{K} p_{kj}^{S}} \tag{13}$$

When

$$(\sum_{j=1}^{C} D_{jk}^{1/1-m})^{1-m}$$

is not large enough, its differential coefficient is much larger than the corresponding differential coefficient of ε (N, ϕ). In this case, we ignore the second section and the corresponding objective function turns out to be

$$J_C^S = \sum_{k=1}^{K}\rho(\sum_{j=1}^{C} D_{jk}^{1/1-m})^{1-m} \tag{14}$$

In view of the convenience, we can think of the $\rho(\cdot)$ as a linear function approximately. Suppose its differential coefficient is the constant Cons and in this case the objective function need to be satisfied:

$$\left\| x_k^S - v_j^S \right\|^2 - \frac{\sum_{\substack{i=1\\ i\neq S}}^{N} p_{kj}^{i}}{p_{kj}^{S}} = 0 \tag{15}$$

Through formula (19), we can derive:

$$p_{kj}^{S} = \frac{\dfrac{\sum_{\substack{i=1\\ i\neq S}}^{N} p_{kj}^{i}}{\left\| x_k^S - v_j^S \right\|^2}}{\sum_{i=1}^{C}\dfrac{\sum_{\substack{l=1\\ l\neq \Gamma}}^{L} p_{ki}^{l}}{\left\| x_k^S - v_i^S \right\|^2}} \tag{16}$$

Formula (10), (13) and (16) compose the cooperative clustering algorithm MICCA and it can be summarized as:

Algorithm MICCA

Input: the initial cluster center $c_1^{(0)}, c_2^{(0)}, \ldots, c_k^{(0)}$, the clustering objective function threshold $CP_{threshold}$

Output: the clusters' center c_1, $c_2, \ldots, c_k$

- Set the initial cluster center be $c_1^{(0)}, c_2^{(0)}, \ldots, c_k^{(0)}$ and the iterative number inum = 0;
- According to formula (10), (16), calculate $J_C^{S(inum)}, p_{kj}^{S(inum)}$;
- According to formula (13), calculate centroid $v_1^{s(inum+1)}, v_2^{s(inum+1)}, \ldots, v_k^{s(inum+1)}$;
- Use the calculated cluster center $v_1^{s(inum+1)}, v_2^{s(inum+1)}, \ldots, v_k^{s(inum+1)}$ in step 3, formula (10) and formula (16), calculate $J_C^{T(inum+1)}, p_{kj}^{T(inum+1)}$;
- If

  $$\left| J_C^{T(inum)} - J_C^{T(inum+1)} \right| < CP_{threshold}$$

  is satisfied, the algorithm stops and $v_1^T, v_2^T, \ldots, v_k^T$ are the final cluster centers in dataset T; otherwise, set to be inum+1 and go to step 3.

RESULTS

In this section, we will test the performance of MICCA algorithm under cooperative circumstances. Experiment results show that cluster center's track of MICCA is different from conventional independent clustering algorithm. And more, the final cluster center is more compact and harmonious.

Now we see about the variation of mutual information measurement defined in formula (11) in clustering process further. From Fig. 1, we can see that with clustering proceeds, cooperative relations among datasets become closer and closer. In more details we can explain this phenomenon as: in initial clustering, large volume of data in datasets is in confused order, no any ordered structure. In this phase, cooperation among datasets shows weak. With clustering proceeds, more and more ordered hidden in the datasets' structure will be. Similarly, in this stage, the cooperative relations are enhanced among different datasets. It is worthwhile to note that in Fig. 1 the cooperative relations are enhanced faster in left area than those in right area. This phenomenon is caused by more steadily structure in left area than in right area.

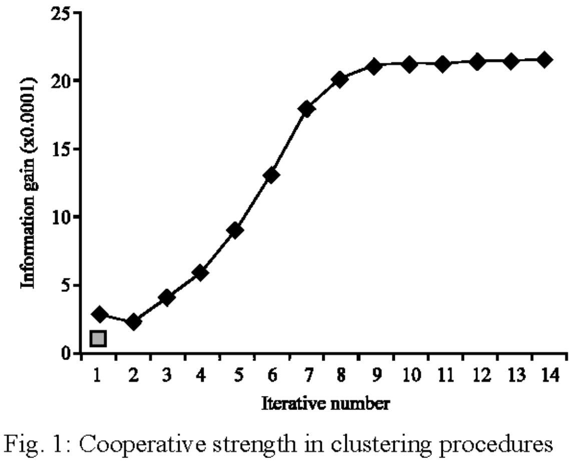

Fig. 1: Cooperative strength in clustering procedures

In fact, all of samples in dataset are affected by MICCA considering cooperative relations hidden in many datasets. Selecting some typical points in dataset, we compare their dependence degree between considering and not considering cooperative effects.

CONCLUSIONS

In real world, a dataset is independent of other datasets but sometimes can be cooperative with others. Conventional clustering algorithms ignore this kind of cooperative relations. In this study, a novel collaborative model is discussed and new proper methods such as mutual information are proposed to quantitatively measure such collaboration between datasets. The corresponding collaborative clustering algorithms are presented accordingly and the theoretic analysis shows that the new cooperative clustering algorithms can finally converges to local minimum. Experimental results demonstrate that the clustering structures obtained by new cooperative algorithms are different from those of conventional algorithms for the consideration of collaboration and the performances of these collaborative clustering algorithms can be much better than those conventional isolated clustering algorithms under the cooperating circumstances.

ACKNOWLEDGEMENT

This work is supported by the National High Technology Research and Development (2001AA113182).

REFERENCES

Han, J. *et al*., 2001. Data Mining: Concept and Techniques. Morgan Kaufman Publishers.

Hopper, F., 1999. Fuzzy Cluster Analysis. Chichester: John Wiley.

Shen, H. *et al*., 2005. Study on new information theory based cooperative clustering algorithm. Chinese J. Computers, 28:1287-1294.

Kanungo, T. *et al*., 2002. An efficient k-means clustering algorithm: analysis and implementation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 24: 881-892.

Topchy, A. *et al*., 2004. A mixture model of clustering ensembles. In: Proceedings of the SIAM International Conference on Data Mining, pp: 22-24.