Journal of Theoretical and Applied Information Technology
20th February 2013. Vol. 48 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

AN IMPROVED SEMI-SUPERVISED CLUSTERING ALGORITHM BASED ON ACTIVE LEARNING

[1]ZHANG CHUN NA, [2]ZHU YONG YONG, [3]LI YI RAN

[1.3] College of Software ,University of Science and Technology Liaoning, Anshan 114054, China

[2] Department of Economics and Business Administration, Chongqing University of Education, Chongqing 400067, China

E-mail: [1]zcn1979@yahoo.com.cn, [2]zhuyongyong@126.com [3]lyr7879@yahoo.com.cn

ABSTRACT

In order to solve the difficult questions such as in the presence of the cluster deviation and high dimensional data processing in traditional semi-supervised clustering algorithm, a semi-supervised clustering algorithm based on active learning was proposed, this algorithm can effectively solve the above two problems. Using active learning strategies in algorithm can obtain a large amount of information of pairwise constraints therefore enhance the proportion of prior knowledge. And the use of this constraint set projection space, finally in the mapping of the subspace, the improved K-means algorithm implemented for data clustering, as the algorithm clustering object is a low dimensional data, and prior knowledge increased, clustering in time efficiency can be guaranteed, and also can solve the deviation problem of clustering. The experiment results show that, with active learning algorithm clustering performance improvement, was superior to the other two semi-supervised clustering algorithms.

Keywords: *Pairwise Constraints; Semi-Supervised Clustering; K-Means Algorithm; Active Learning*

1. INTRODUCTION

In the process of solving the practical problems by using data mining, we often encounter some cannot be labeled data. If using artificial markers, it is too costly on the one hand and on the other hand it will cause unexpected damage easily. Therefore, how to use limited prior knowledge, which comes from the data related to the small category labels and constraint condition to complete clustering analysis has become a hot issue in recent years.

At present, semi-supervised clustering algorithm can be divided into three categories: The first category is based on the constraint of semi-supervised clustering algorithm, derived from the pairwise constraints proposed by Wagstaff et al: must-link and cannot-link[1]. These algorithms are determined in dependence on the above two kinds of constraints, results of the two constraint are the opposite. Among them, must-link provided that two data samples in the space belong to the must-link constraint, then the two divisions as a class; on the contrary, cannot-link provided that two data samples in the space belong to the cannot-link constraint, the two data are divided into different classes. The second category is based on the distance of the semi-supervised clustering algorithm and using trained adaptive distance metric to evaluate, through movement of the sample produced different distances, which constructed the restriction conditions to meet clustering [2]. In addition, two kinds of algorithm can also be combined to implement clustering[3], it is the so-called third category.

These three semi-supervised clustering algorithms have several common problems: first, cluster deviation, using pairwise constraints in must-link and cannot-link study, sample points around a cluster center move now and then in order to obtain the best position, the distance of sample points in the algorithm iterations is changing. Note, must-link does not guarantee that all the corresponding constraint sample point is divided into one class and also cannot-link constraint cannot guarantee that it can be classified into different categories, there exist certain errors; Second, supervisory information is usually non-active ways of obtaining in semi-supervised clustering, the collection of all possible supervisory information is obviously not feasible by traversing, therefore only under limited conditions can obtain some valuable information. Because of pairwise constraints semi-supervised clustering algorithm limitations, it is often too small that information embodied in the

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

constraint set, then influence the overall effect of clustering. In addition, the sample space is high dimensional samples, and the spacing between sample points has smaller difference, the algorithm processing ability is also poor. So how to minimize the cost reduction is a research focus.

In view of the above problems, this paper puts forward a kind of semi-supervised K-means clustering algorithm based on active learning, to obtained the projection matrix under the action of the pairwise constraints and implemented LDA(Linear Discriminant Analysis) reduce dimension to it, while taking advantage of the K-means algorithm to guide the clustering. The method not only solves the problem of clustering deviation, but also play the role of dimensionality reduction, and reduce the computational complexity, improve the clustering performance. Furthermore, joining idea of active learning in algorithm and extracting actively supervision information used feedback in the clustering process can solve the issue that constraint set relates to the amount of information is too small and make the clustering effect is much better. The experiment data proved that algorithm is efficient and the result of clustering is satisfactory.

2. RELATED KNOWLEDGE AND RESEARCH

2.1 Semi-Supervised Clustering

Based on the above viewpoints, semi-supervised clustering research can be roughly divided into three directions: Based on the constraint mechanism, based on the distance and hybrid. Related research at present basically belong to the three class, which based on the pairwise constraints algorithm include: reference[4] is based on density clustering algorithm, can deal with any shapes of clusters, and based on the constraint set to split or merge clusters; reference[5] presented an effective semi-supervised clustering algorithm and introduced fuzzy constraint thought, with minimal supervision information clustering; reference[6] puts forward a kind of distinguishing nonlinear transformation metrics in measurement and based on image retrieval to test , its effect is good.

2.2 Active Learning Algorithm

Active learning algorithm is a branch of classification algorithm, because of the relatively wide research direction and application, domestic and foreign scholars have put forward many topics. Reference[7] use source domain data to study the target domain with active learning algorithm, trying to simplify the sample point label complexity. In reference[8] Tomanek et al described the important application of active learning in the NLP (Natural Language Processing), focus on how to create high-quality training sample set. Ambati et al analyzed word alignment model in machine translation system, which helps to reduce the data word alignment error rate by creating the half word alignment model combining unsupervised and supervised learning, and makes data concentration abnormal or makes noise sensitive[9].

3. ALGORITHM ANALYSIS

First, this section proposed semi-supervised document clustering algorithm based on pairwise constraints, algorithm introduced pairwise constraints in the K-means[10], the use of LDA redefined cluster space to carry out the process of clustering at the same time, and then through the active learning algorithm to obtain more supervision information, to improve the performance of the algorithm.

Set finite sets in D dimension space Q^D $X=\{x_1,x_2,\cdots,x_n \mid x_i \in Q^D\}$, define S_m as must-link pairwise constraint set， $S_m=\{x_i,x_j\}$; define S_c as cannot-link pairwise constraint set, $S_c=\{x_m,x_n\}$. Implementation process of semi-supervised K-means clustering algorithm based on active learning can be divided into three steps:

Step 1: The initialization of algorithm, using active learning algorithm for a given must-link and cannot-link pairwise constraints set for processing, in order to get abundant information of pairwise constraints, and then obtain the corresponding projection matrix;

Step 2: Mark the vector of the projection matrix and use LDA redefine cluser space;

step3: Implement clustering to the training set by using K-means algorithm.

3.1 Algorithm Initialization

Active learning algorithm plays the role of perfecting pair constraint set, and makes information contained in the constraint set as much as possible, in order to improve the performance of clustering.

For an irregular cluster, constraint set that mature algorithms need shall have the following two conditions:

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

(1)At least one non-specific elements should be contained in the constraints set in clustering;

(2)Each cluster boundary should be controlled by the corresponding pairwise constraints.

Based on this, this paper designs a kind of semi-heuristic active learning algorithm, supplemented by a variety of preset parameters to achieve the set of constraints of information maximization. The algorithm flow is as follows:

Algorithm 1: Semi-heuristic active learning algorithm.

Input: Sample document A, pairwise constraints number N, the radius of core region of r, core area sample threshold ε.

Output: Pairwise constraints set.

Step 1: Initial sample document A, use given r and ε to establish core region of sample point, marked as C_A, and the boundary point set, marked as $_{B_A}$;

Step 2: Pairwise constraints set S is initialized as the empty set;

Step 3: With the precondition of not cross the line of pairwise constraints set, using model respectively determine the core region and boundary point sample classes, to build on the pairwise constraint set $\{S \mid x, y\}$.

Pseudo code:

```
fn_InitDoc(A);
// Initialization of the sample document
Public ObjPaCons fn_Activlearn(int N,double d_Cradius,String s_Threshold,objRange obj_CA, objRange obj_BA)
// build on the pairwise constraint set
{ int i_ConsCount=0;
  ObjPaCons obj_cons=new ObjPaCons();
  // Initialization of the pairwise constraint set
  ObjRange obj_CAS;
  fn_InitCA(obj_CAS);
   // Initialization of core zone point set
  while(i_ ConsCount< N)
  //Judge the obtained constraint number
 {// through the judgment of core set, establish
     addition of various locations
     if (obj_CAS is NULL)
      fn_InsertCA(obj_CA,obj_CAS,N);
     else
      fn_InsertCA(obj_CA,obj_CAS,F);
      fn_ConstrucCons(obj_CAS);
    // Traverse the core set, build on pairwise constraints set
        … …
    //Initialize the boundary set and construct the corresponding pairwise constraints set
  }
  return obj_cons;
}
```

In order to ensure that each class has an element contained in a pairwise constraints, this paper adopts the most far priority strategy to select the core point set, the distance between a core point x and the core point set obj_CAS is the minimum cosine distance under the condition of low enough attempt, so the cluster center can be more than one in clustering, method is reasonable and efficient, and without loss of generality.

In the initialization of boundary set, choose two points from core A nearest and furthest respectively as the boundary point to construct the pairwise constraints set.

3.2 Establish Projection Matrix

When building pairwise constraints set we focus on the problem of core point set and the boundary set, here the distances between points in space is usually of low dimension, high dimensional point is unable to measure distance, or the distance is the same. Therefore, how high dimensional spatial data are mapped to a lower dimensional space is a problem that must be considered, with the vector of the projection matrix $M_{l\times k} = \{m_1, m_2 \cdots, m_k\}$, in which every vector is of dimension l orthogonal unit vectors, elements of x_i projected onto the low-dimensional space is given as follows:

$$x^i = M^T x^i, \quad l < k \tag{1}$$

For the transformed data, the points that according to the characteristics of cannot-link constraints corresponding to should be kept as far as possible the most distance, in contrast, the points that must-link constraints corresponding should be maintained a close distance. Therefore, the projection matrix construction principle embodied in a data structure projection consistency. Here, using the objective function $P(M)$ to complete the data transform.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

$$P(M)=\frac{1}{2}\sum_{i,j}\left\|M^T x_i-M^T x_j\right\|^2 + \frac{\alpha}{2\left|C_n\right|^2}\sum_{(x_i,x_j)\in C_n}\left\|M^T x_i-M^T x_j\right\|^2 - \frac{\beta}{2\left|M_n\right|^2}\sum_{(x_i,x_j)\in M_n}\left\|M^T x_i-M^T x_j\right\|^2 \quad (2)$$

In the formula, C_n , M_n were pairwise constraints number of cannot-link and must-link respectively; Considering the constraint factors, setting balance coefficient α 、 β , so as to adjust the proportion of the target function.

The simplified formula (2):

$$P(M)=\frac{1}{2}\sum_{i,j}\left\|M^T x_i-M^T x_j\right\|^2 Q \quad (3)$$

In the formula,

$$Q=\begin{cases}1+\dfrac{\alpha}{\left|C_n\right|^2} & x_i,x_j\in C_n\\ 1-\dfrac{\beta}{\left|M_n\right|^2} & x_i,x_j\in M_n\\ 1 & x_i,x_j\notin C_n \text{ or } M_n\end{cases} \quad (4)$$

Two derivation of formula (3):

$$P(M)=M_i^T M_j\left(\frac{1}{2}\sum_{i,j}\left\|x_i-x_j\right\|^2 Q\right) \quad (5)$$

In the formula, $M_i^T M_j = i\begin{cases}1 & i=j\\ 0 & otherwise\end{cases}$

Through derivation of formula (3), the optimized matrix M can be got. Using Lagrange theorem, analytic solution of matrix equation M to the target and constraints can be obtained, formula is as follows:

$$R(M)=P(M_1,M_1,\cdots M_n)-\sum_{j=1}^{n}\gamma_i(W_i^T W_j-1) \quad (6)$$

Put formula (6) and seek partial derivative of $R(M):\frac{\partial R}{\partial M_j}=QM_j-\gamma_i M_i=0$, it can solve the optimal projection matrix vector $M_{m\times n}=\{M_1,M_2,\cdots,M_n\}$ and use formula (1) to transform high dimensional data space into a low dimensional data space, so as to realize clustering algorithm.

In cluster, from the processing difficulty, low dimensional data is undoubtedly the best choice. By the formula(2) and cannot-link , must-link two pairwise constraints, calculate the projection matrix, we can obtain low dimension space data under the premise of maintaining the original data structure. Here, should control the distance of cannot-link constraint point set as large as possible, that of the must-link constraint set as small as possible, so that the cluster effect is better.

3.3 Cluster Deviation Analysis

For the deviation of clustering, the traditional solution is added balance factor in the sample space, make cannot-link and must-link effect really, sample point balance type as follows:

$$\phi_{ij}=\frac{1}{2}\left\|x_i-x_j\right\|^2_{C_{l_i}}+\frac{1}{2}\left\|x_i-x_j\right\|^2_{C_{l_j}},\qquad \overline{\phi_{ij}}=\frac{1}{2}\left\|x'_{l_i}-x''_{l_j}\right\|^2_{C_{l_j}}-\frac{1}{2}\left\|x_i-x_j\right\|^2_{C_{l_i}} \quad (7)$$

For the two balance factor is described: if the sample x_i , x_j violates the constraint of cannot-link, ϕ_{ij} stands for summation of two samples in the difference between the distances under different metric system; If the sample x_i , x_j violates the constraint of must-link, $\overline{\phi_{ij}}$ is the difference value of distance between two samples of longest distance and current sample point in clustering. Using the balance of factors can reduce cluster deviation in certain extent, but it is just a simple adjustment, the effect is not obvious.

Based on the above analysis, this paper put forward cannot-link that using the adjusted virtual sample points instead of actual sample point constraint to solve the clustering deviation, concrete analysis is as follows:

Definition 1: The same cluster closure, sample points set $\{x_1,x_2,\cdots,x_n\}$, in the formula, $(x_i,x_j)\hat{I}\ S_m, i^1\ j, 1\pounds i,j\pounds n$, then the set formed by $\{x_1,x_2,\cdots,x_n\}$ is called the same cluster closure.

Definition 2: special cluster closure, supposed there are two special closure sets $X=\{x_1,x_2,\cdots x_n\}$ and $Y=\{y_1,y_2,\cdots,y_m\}$, in the formula, $(x_k,y_l)\hat{I}\ S_c, 1\pounds k\pounds n, 1\pounds l\pounds m$, and $x_k\hat{I}\ X$, $y_l\hat{I}\ Y$, then X and Y are closures.

Journal of Theoretical and Applied Information Technology
20th February 2013. Vol. 48 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Definition 3: closure centre, supposed there exists the same cluster closure set $\{x_1, x_2, \cdots, x_n\}$, then define $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ as the centre of the closure set.

Simplify sample set according to the definition of closure and attempt to use the center to replace the original sample closure between cannot-link constraints. Fig 1 use hollow circle represents the sample point, solid circle represents the sample set closure center, the solid line shows the sample points between must-link constraints, and the dotted line represents the sample points between cannot-link constraints. Two sample sets $\{x_1, x_i, x_{i+1}, \cdots, x_n\}\{y_1, y_j, y_{j+1}, \cdots, y_m\}$ represent two closures respectively, thus the cannot-link constraints between two samples sets can be replaced by closure center $\bar{x}, \bar{y}$ .There is a special case, if a sample is not subordinate, the individually into a closure, the sample set contains only one sample point.

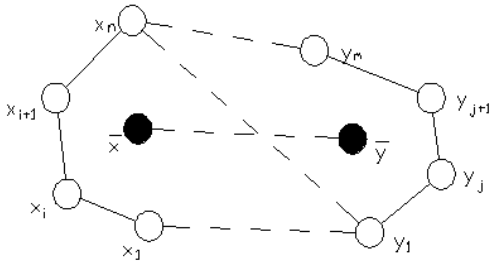


Figure 1: Diagram of sample set closure

Replace sample set closure with the closure center, cannot-link constraint in special cluster closure take the place of in closure center, the size of the sample set is greatly reduced, can be transformed into $X = \{\bar{x}_1, \bar{x}_2, \cdots, \bar{x}_l\}, l \leq n$, cannot-link constraint set of sample points are $S_c = \{\bar{x}_m, \bar{x}_n\}$.

To be clear, the clustering objects have been changed into closure from sample points. The closure center and the cluster center are very close physically and logically. This paper attempts to analyze those results of alternative closures after clustering without deviation. Set $X_i = \{x_1, x_2, \cdots, x_l\}$ as same cluster closure in sample set, closure centre as $\bar{x}_i$, k cluster centre can be represented as $Z = \{z_1, z_2, \cdots z_k\}$, k cluster can be represented as $\{c_1, c_2, \cdots, c_k\}$. $\forall z_i \in Z$, let $\arg\min(\| \bar{x}_i - z_i \|^2) \leq \varepsilon$, therefore $\bar{x}_i \in C$. $\bar{x}_i$ replaced X_i , $\bar{x}_i \in C_i \Rightarrow X_i \in C_j, \bar{y}_j \in C_j \Rightarrow Y_j \in C_j$. So, after replacing, the same cluster closure X_i is still vested in the original cluster, satisfying must-link. The above validation for different cluster closure is also applicable in special clusters, closure center were $\bar{x}_i, \bar{y}_j$, X_i and Y_j are different cluster closure, and each is the same cluster closure, all two meet the must-link constraint, that is $\bar{x}_i \in C_i \Rightarrow X_i \in C_j, \bar{y}_j \in C_j \Rightarrow Y_j \in C_j$. So, to ensure the mutually different cluster closure of X_i, Y_j belong to different categories, to meet the cannot-link constraints, therefore, using sample closure can solve the clustering subject.

3.4 Improved K-Means Algorithm

Algorithm 2: Improved K-means algorithm based on pairwise constraints.

Input: sample set X_l , algorithm cluster number k , cannot-link constraints set S_c .

Output: k Clusters division.

Step 1: Calculation k clusters centre in sample set;

Step 2: Calculated $\arg\min$ that the center of closure and clustering among the same cluster and special closure respectively, iterative process go round and begin again, until the algorithm convergence;

While (convergence conditions)

(1)For one same cluster closure, calculate closure center $\bar{x}_i$ and its $\arg\min(\| \bar{x}_i - z_i \|^2)$ of cluster centre z_i , so that it will fit conditions;

(2)For special closure, two closure centre $\bar{x}_i$ and $\bar{y}_j$, there is $(\bar{x}_i, \bar{y}_j) \in S_c$, calculate $\arg\min(\| \bar{x}_i - z_i \|^2 + \| \bar{y}_j - z_j \|^2)$ of two cluster centre with it;

(3)Calculate the cluster centre z_i .

Step 3: Return k divided cluster.

To solve the clustering problems, the traditional approach is introduced the balance factor to regulate the constraint violation sample point distance value, but the balance factor values for different objects to determine is difficult, the effect

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

is not good. This paper proposes the use of sample closure center instead of sample closure to meet cannot-link and must-link constraints, and in the algorithm K-means realize, due to alternative that is involved in computing the sample points are greatly reduced, the efficiency of the algorithm can be ensured, suitable for complex sample set.

4. EXPERIMENT

4.1 Experiment Contents

This paper selects two data sets from the UCI database: Balance, Similar. Among them, balance is low dimensional data set, the dimension is 4, sample number 585; similar is high dimensional data set, the dimension is 16090, sample number 280. Use two quantitative indexes: NMI(normalized mutual information) and PCM(pairwise comprehensive measure). NMI is a kind of clustering effect evaluation index, response samples clustering results and real class similarity, its range is set to [0,1], the larger the better description of clustering; PCM combined precision and recall rate, its range is set to [0,1], same as previous, the larger the better clustering effect.

This paper involved three kinds of algorithm: K-means algorithm based on pairwise constraints, mark PC-KMS; pairwise constraints based on improved K-means algorithm, denoted as CPC-KM, as well as the text presents the active algorithm CPCKMEANS, mark ACPC-KM. The experiment was divided into three parts:

(1)Compare time efficiency of PC-KM, CPC-KM and ACPC-KM;

(2)Compare cluster effect of PC-KM、CPC-KM and ACPC-KM;

(3)Analyze NMI value of PC-KM、CPC-KM and ACPC-KM.

In PC-KMS, CPC-KM and ACPC-KM performance comparison of the three algorithms, each algorithm run 20 times on the data set, the final clustering results take the average value of 20 times. In addition, the set of clustering number k and data set of true category should be coordinated, and descend dimension data set to $k-1$, Each experiment of constraint sets include cannot-link constraints and must-link constraints, and the number of constraint consistent, Fig 2 is the comparison of three kinds of algorithm in clustering effect; Fig 3 is the comparison of NMI value among PC-KMS, CPC-KM and ACPC-KM.

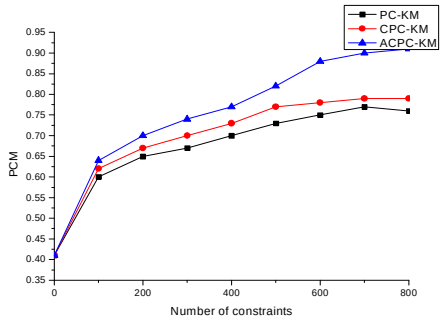
4.2 Result Analysis

The figure respectively reflect comparison results of algorithm in paired comprehensive measure and normalized mutual information, when the constraint number is 0, three kinds of algorithms and results are the same because of no room for improvement.

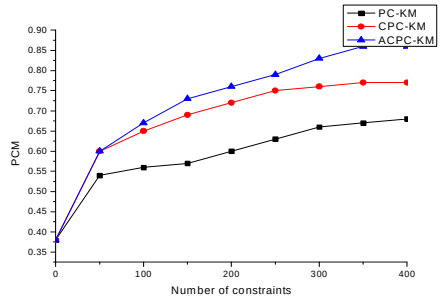
Experiment results of comparison of three algorithms in paired comprehensive is shown in Fig 2. Two data sets are verified for clustering effects from low to high arrangement: PC-KM, CPC-KM, ACPC-KM. In low dimensional data sets Balance, PC-KM and CPC-KM algorithm differs not quite because there is no need to consider the effects of dimension, but in the CPC-KM algorithm with closure center instead of sample points to meet cannot-link and must-link constraints, the algorithm has some advantages. In the high dimensional data sets Similar, because PC-KM inherent spatial processing capacity leads to dividing the sample error increase. therefore, CPC-KM and ACPC-KM algorithm have more obvious advantages. In addition, with the increase of the number of ACPC-KM constraints, the rising slope change, also confirmed the importance of constraint set information quantity. Fig 3 mainly embody the NMI index of algorithm, three kinds of clustering performance of the algorithm with the constraint number increases gradually, but the different constraint number corresponding to the clustering performance is different, in the Balance data sets, when the number of clusters in 600-800, ACPC-KM algorithm NMI values based on active learning increased significantly, whereas the other two algorithms for clustering performance declined, because PC-KM and CPC-KM algorithm of pairwise constraints of randomly generated, the result shows the active learning effect.

From the experimental results, compared with the PC-KM and CPC-KM algorithms, ACPC-KM clustering effect is more obvious. Whether it is a low dimensional or multidimensional data, the number of pairwise constraints at higher inflation rates, because the ACPC-KM algorithm constraint set contains more information so that the ability that control sample boundary is powerful.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

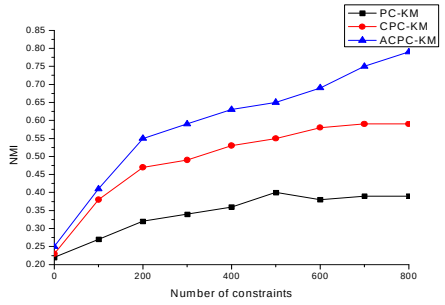


（a）Balance

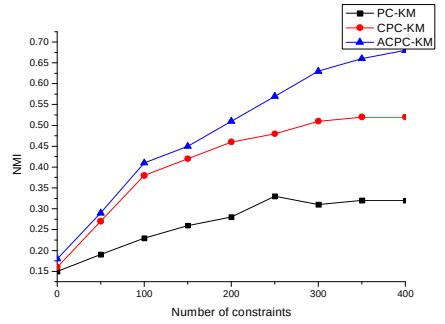


（b）Similar

Figure 2: Comparison Result Of PCM Quota



（a）Balance



（b）Similar

Figure 3: Comparison Result Of NMI Quota

5. CONCLUDING REMARKS

This paper presents a semi-supervised clustering algorithm (ACPC-KM algorithm) based on active learning, which using cannot-link constraints and must-link constraints to guide the clustering, in order to increase the constraint set information quantity, introduced active learning strategies, each iteration obtain a large number of auxiliary information, at the same time through pairwise constraints obtain initial projection matrix, and construct the subspace. Finally, using the improved K-means algorithm to implement clustering sample set. The experimental results show that the ACPC-KM algorithm not only has the time efficiency and solve the problems in clustering has obvious advantages, can effectively improve the performance of clustering.

In semi-supervised clustering, more prior knowledge, better cluster effect, but the supervision information will add too much burden on the user, reasonable cannot-link and must-link constraints can be more effective results, therefore, how to improve the constraint set information quantity is an important research direction.

REFRENCES:

[1] Wagstaff K and Cardie C, “Clustering with instance-level constraints”, *Proceedings of the Seventeenth International Conference on Machine Learning*, 2000:1103-1110.

[2] Wagstaff K, Cardie C, and Rogers S, et al, “Constrained K-Means Clustering with background knowledge”, *Proceedings of the Eighteenth International Conference on Machine Learning*, 2001:577-584.

[3] Bilenko M, Basu S, and Mooney R J.Integrating, “Constraints and Metric Learning in Semi-Supervised Clustering”, *Proceedings of the 21st International Conference on Machine Learning*, 2004:81-88.

[4] Ruiz C, Spiliopoulou M, and Menasalvas E, “Density-based semi-supervised clustering”, *Data Mining and Knowledge Discovery*, 2010,3(21):345-370.

[5] Ioannis A. Maraziotis, “A semi-supervised fuzzy clustering algorithm applied to gene expression data”, *Pattern Recognition*, 2012,1(45): 637-648.

[6] Hong Chang and Dit-Yan Yeung, “Locally linear metric adaptation with application to semi-supervised clustering and image retrieval”, *Pattern Recognition*, 2006,7(39):1253-1264.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[7] Rai P, Saha A, Hal Daumé, III, Venkatasubramanian S, "Domain adaptation meets active learning", *ALNLP '10 :Proceedings of the NAACL HLT 2010 Workshop on Active Learning for Natural Language Processing*, 2010,27-32.

[8] Tomanek K, Olsson F, "A web survey on the use of active learning to support annotation of text data", *HLT '09:Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing*, 45-48.

[9] Ambati V, Vogel S, and Carbonell J, "Active semi-supervised learning for improving word alignment", *ALNLP '10: Proceedings of the NAACL HLT 2010 Workshop on Active Learning for Natural Language Processing*, 2010,10-17.

[10] Ares ME, Javier Parapar, and Barreiro Álvaro, "An experimental study of constrained clustering effectiveness in presence of erroneous constraints", *Information Processing and Management*, 2012,3(48): 537-551.