Journal of Theoretical and Applied Information Technology
20th March 2013. Vol. 49 No.2

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

THE CLASSIFICATION OF IRREGULAR CLUSTERS: AN IMPROVED SEMI-SUPERVISED CLUSTERING ALGORITHM

[1.2]MA CHI, [1]GAO XUEDONG, [2]ZHANG CHUNNA, [2]LI YIRAN

[1]Dongling School of Economics and Management, University of Science and Technology Beijing, Beijing 10083, China

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

E-mail: asmachi@126.com

ABSTRACT

This paper proposed a improved semi-supervised clustering algorithm, aiming at the limitations of the traditional k-means algorithm in the irregular clusters division. This paper mainly discusses two problems: the closure replacement and the traction between data points. First of all, made use of the closure center to replace the original sample point, and to calculated the sample clustering center combining the semi-supervised clustering ideology with a small amount of the tagged data; next, in the process of clustering is introduced into the definition about the traction distance of the clustering center and the traction between the data points, given a complete description of the relationship between data points; finally, fully consider the effect about the tagged data points, especially isolated point, on untagged data point, implement clustering data in the sample space. Experimental results indicate that the cluster is much more efficient than the other existing algorithms when dealing with the irregular cluster.

Keywords: *Semi-supervised clustering, K-means algorithm, Closure Center, Irregular cluster*

1. INTRODUCTION

The cluster analysis is the most important research directions of the knowledge discovery and data mining areas, in accordance with certain rules, which is divided into different classes, the higher similarity of data is divided into a class, and the lower divided into another class. With the deepening of the data mining technology applications in various fields, the clustering analysis is one of the most frequently used technique, which also more and more is converted into a practical application.

The semi-supervised clustering is an emerging cluster analysis method, which combines the traditional supervised learning and the unsupervised learning[1], to make use of a priori knowledge supports clustering, its advantages are the completed classification in the case of less knowledge, the supervised learning object can be both the tagged data and untagged data, so that the clustering process will become very easy.

In 1967, the k-means algorithm was first proposed by MacQueen, which has gradually developed one of the most commonly used cluster analysis algorithm[2]. With the development of the semi-supervised clustering, the research which combined two becomes more and more [3-5].

Making use of the underlying distance metrics, the reference [6] designed a fast k-means algorithm to annotation the relationship between the figures. In comparison, the k-means algorithm is not much used in semi-supervised clustering. The reference [7] proposed a kind of semi-supervised clustering algorithm based on kernel, the clustering process adopt the mode of c-means combined with k-means, the algorithm accuracy rate is very high, but the complexity of high dimensional data is also higher. The reference [8] provided the semi-supervised clustering algorithm for pattern discovery, the information is extracted from the training set, using active learning to find more attributes, taking the initiative to reduce attributes of the regular expression. Because the attribute is unknown and each test requires a large amount of contrast, the algorithm is time-consuming. The reference [9] presented a structure of assemblage based on the density stratification about the expansion of the tagged object. The algorithm in the paper is better in efficiency and quality than the original algorithm to get assemblage, especially to treat the data on the distribution of multi-peak within an association. In addition, some algorithm based on the pairwise constraints or the active learning, and so on[10,11]. This paper aims at that the k-means algorithm for the limitation of irregular clusters, put forward a kind of improved semi

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

supervised clustering algorithm, which using sample point closure center instead of dispersed sample point calculation of cluster centers, and introduce the traction force between data points to build traction matrix, constantly optimize the cluster centers in process of iterative, in order to seek the optimal solution. The experimental part of this paper analyses and contrasts the clustering accuracy of different clustering algorithms. The experiments show that the improved semi-supervised k-means algorithm clustering effect is better.

2. THE SEMI-SUPERVISED K-MEANS CLUSTERING ALGORITHM

2.1 The k-means algorithm

Set the finite set in Q dimensional space S^Q as $X=\{x_1,x_2,\cdots,x_n\}$, the initialization can be divided into k class randomly, denoted as $C_1,C_2,\cdots,C_k$, if a class has n objects, the i clustering center can be defined as $Z_1,Z_2,\cdots,Z_k$, $Z_i=\frac{1}{n}\sum_{j=1}^{n}x_j, j\in[1,k]$, the definition of the objective function as follows:

$$J=\sum_{i=1}^{k}\sum_{j=1}^{n_j}D^2_{x_j,z_i} \quad (1)$$

In the formula, $D^2_{x_j,z_i}$ represents the distance that the j text to a class i clustering center, that is, Euclidean distance.

The core idea of this algorithm is through continuous iteration to find the k optimal cluster centers of the sample data set, and other data move to the cluster center, until the objective function value is minimum.

2.2 The algorithmic thought

The semi-supervised clustering sample data sets can be classified as follows: the tagged data set and the untagged data set, there are nL tagged data in the training set, the tagged data result set can be expressed as: $x_i=\{x_1,x_2,\cdots,x_{nL}\,|\,1\le i\le nL\}$, the remaining sample data is the untagged data, it is expressed as: $x_j=\{x_{nL+1},x_{nL+2},\cdots,x_n\,|\,nL+1\le j\le n\}$. In the formula, $x_i,x_j\in X$, and $x_i\cup x_j=X$. The tagged data in semi-supervised clustering usually have three kinds of functions: the clustering seed, the clustering limit and the clustering feedback. This article discusses the data graphically showed a cluster of irregular and dynamic, namely the clustering previous number of entries is not known. Based on this kind of uncertainty should be considered the third function of the semi-supervised clustering, namely the clustering feedback.

Through the use of tagged data as the clustering feedback for semi-supervised clustering, first of all, organizing the training set and using the unsupervised of spherical K-means algorithm implemented clustering; next, the result of clustering is introduced the tagged data, and to fine-tune it, the adjustment results as the initial value of the again cluster; finally, noting the iteration is needed to keep the cluster number unchanged, at the same time detection value of the objective function is or not to achieve the desired value, if fulfilled, is terminated, or to continue. The algorithm contains the following steps:

Input: The sample data set $X=\{x_i\,|\,1\le i\le n\}$, and $x_i\in S^Q$, define k clusters randomly, the relationship between the tagged data sets with the sample data sets expressed as: $X_1\cup X_2\cup\cdots\cup X_k\cup X_j=X$.

Output: The clustering results are the objective function of the minimum k divided set: $\{X_m\,|_{m=1}^{k}\}$.

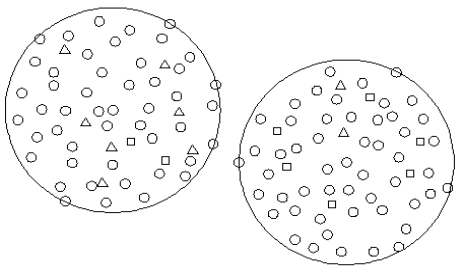
Step 1: Using the tagged data to calculate the initial cluster center: $Z_m^0\,|_{m=1}^{k}$, to make $Z_m^i=\frac{\sum x_i}{X_m}\,|\,x_i\in X_m$, Set the iteration operator $l=0$;

Step 2: Assigned the data of X to the similar clusters, referring the initial cluster centers;

Step 3: The cluster centers are calculated according to the formula in step 1 again, and updated iteration operator l, to make $l=l+1$, the process constantly back and forth until convergence.

2.3 The effect analysis

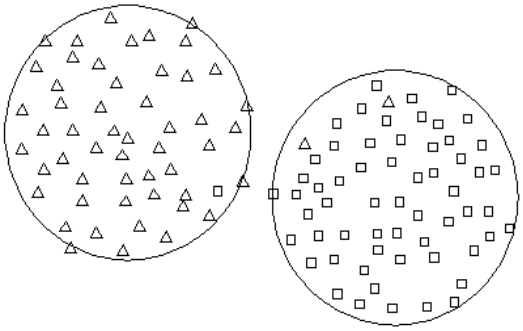
The above discussion algorithm is usually applied to the rule clusters, but the irregular clusters found ability is poor. This article defined it as KM algorithm.



A. The Spherical Cluster Before

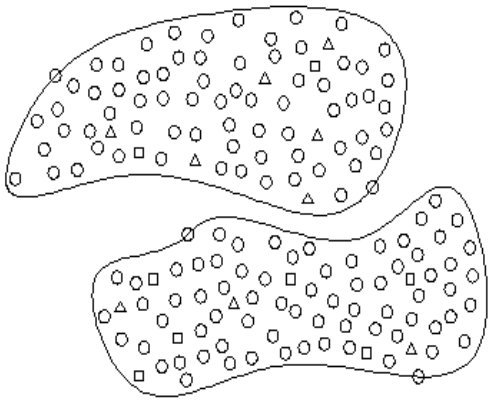
Journal of Theoretical and Applied Information Technology
20th March 2013. Vol. 49 No.2

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

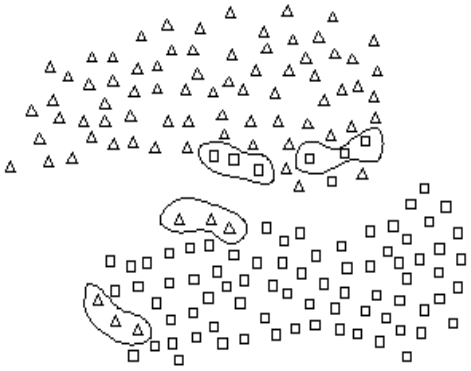


B. The Spherical Cluster After

Figure 1: The Comparison Chart Of Before And After About Spherical Cluster



A. The Irregularly Shaped Cluster Before



B. The Irregularly Shaped Cluster After

Figure 2: The Comparison Chart Of Before And After About Irregularly Shaped Cluster

In the figure 1 and figure 2, □ represents the untagged data, □ and □ represents the tagged data, the data as a whole is divided into two categories, namely□ and □ . Through analysis, the figure 1 is a regular spherical, after cluster, the division of the cluster is more reasonable, the upper part and the lower part of the figure 1b respectively only one point appears deviation; the figure 2 is an irregular pattern, after the algorithm cluster as showed in figure2b, the upper part and the lower part of the figure respectively six point appear deviation. It can clearly be seen that the algorithm for the clustering effect of regular and irregular graphics are quite different.

Through the analysis of the misjudgment point knows, when the untagged data is wrong divided, there will be a few isolated □ or □ within a certain distance of the nearby, However in other non-isolated point are fewer errors. This can be understood as follows: untagged data on its short distance, there is a convergence of the tagged data, and the further away from the tagged data have been exclusive. In addition, due to the limitation of the traditional K-means algorithm, the effect is poor for irregular-shaped clusters clustering.

3. K-MEANS CLUSTERING BASED ON THE SEMI-SUPERVISED OF THE CLOSURE REPLACEMENT

3.1 The Algorithmic Thought

There is stability in the processing of the conventional semi-supervised clustering algorithm for the irregular clusters treatment, the reason mainly comes from the impact of the tagged data to untagged data. This paper presents an improved K-means clustering algorithm based on the semi-supervised, it is introduced the nearest neighbor point impact factor on the basis of the RKM to improve stability, while taking advantage of the center of the closure instead of the sample point implement clustering to avoid clustering deviation, it needs to consider the following two problems:

(1) The tagged data on adjacent data has effect. The experimentally measured that, in terms of distance to one untagged data, the tagged data of the shorter distance to its influence are far greater than the longer distance, therefore the initially determined by the distance that the untagged data belong to which kinds of tagged data group.

(2)The traction between different kind of tagged data. A group includes different kinds of tagged data and a number of clustering center, when untagged data are dragged by the relevant cluster center, if tagged data are relatively few and the distance to arrive untagged data point is nearly around it, the traction force of the data greater than other types of data.

According to the above, the algorithm needs to consider three impact factors, namely the clustering center, heterogeneous as well as a close neighbor points.

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

Definition 1: Clustering center traction distance: set the sample data set $X=\{x_i \mid 1\le i\le n\}$,and $x_i \in S^Q$, the tagged data set $P=\{x_m \mid_{m=1}^{k}\}$, corresponding to the cluster center $Z_m \mid_{m=1}^{k}$,the cluster distance of data points x_i is defined as: $D_{x_i}=\mid x_i - Z_i \mid$.

Definition 2: Traction between the data points: sample data sets X is divided into k clusters in the beginning, set x_i and x_j belong C_i cluster and C_j cluster respectively, and the tagged sequence sets $P \cup U = X$, among them, U is an untagged sequence sets, traction of the data points of x_j to x_i can be defined as follows:

$$TF(x_i, x_j) = \sum_{x_j \in U} \frac{\lambda}{D_{x_i}^{\ 2}}, x_i \notin U \qquad (2)$$

In the formula, $\begin{cases} 0 & x_i \in U, \text{但} x_i \notin C_j \\ 1 & x_i \in U, \text{且} x_i \in C_j \end{cases}$, λ as a gravity factor, if λ=0 ,then $TF(x_i, x_j)$=0 , there was no difference with the RKM algorithm.

In order to effectively improve the clustering effect, this paper is based on a sample point distance relationship, put forward by using the adjusted virtual sample points instead of actual sample point to solve the problem of large deviation of sample clustering, defined as follows:

Definition 3: The same cluster closure, sample points set $\{x_1, x_2, \cdots, x_n\}$, in the formula, $(x_i, x_j) \hat{I} S_m, i \neq j, 1 £ i, j £ n$, then the set formed by $\{x_1, x_2, \cdots, x_n\}$ is called the same cluster closure.

Definition 4: Special cluster closure, suppose 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, $(x_k, y_l) \hat{I} S_c, 1 £ k £ n, 1 £ l £ m$, and $x_k \hat{I} X$, $y_l \hat{I} Y$, then X and Y are closures.

Definition 5: Closure centre, suppose 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.

With the closure center instead of sample set closure, not only can reduce the size of the sample set, but also can effectively eliminate the effect from the isolated point traction, the sample set transform into $X = \{\bar{x}_1, \bar{x}_2, \cdots, \bar{x}_l\}, l £ n$.

It is noted that the clustering objects becomes the center of the closure from the sample point, here, the closure center and the cluster center, both physical and logic are very close, the following, this paper attempts to analyze, after the closure center was replaced, the clustering results without deviation, set $X_i = \{x_1, x_2, \cdots, x_l\}$ is the same cluster closure in the same sets, the closure centre is $\bar{x}_i$, k cluster centers can be expressed as: $Z=\{z_1, z_2, \cdots z_k\}$, k clustering cluster is expressed as: $\{c_1, c_2, \cdots, c_k\}$. $\forall z_i \in Z$, to make $\arg\min(\| \bar{x}_i - z_i \|^2) \le \varepsilon$, so $\bar{x}_i \in C$. Here, $\bar{x}_i$ instead of X_i , therefore $\bar{x}_i \in C_i \Rightarrow X_i \in C_j, \bar{y}_j \in C_j \Rightarrow Y_j \in C_j$. After the closure is replaced, the same cluster closure X_i still belongs to the original cluster, the above validation also apply equally to different clusters closure. In Special cluster closure, the closures centers respectively are $\bar{x}_i, \bar{y}_j$, X_i and Y_j are different clusters closure each other, and each is in the same cluster closure, namely $\bar{x}_i \in C_i \Rightarrow X_i \in C_j, \bar{y}_j \in C_j \Rightarrow Y_j \in C_j$. In this way, the mutually different clusters closure X_i, Y_j guaranteed belong to a different class, therefore, the use of samples closure can solve the deviation problem of clustering.

3.2 Algorithm steps

By analysis, the traction force in Formula (2) acting on between the data points, the untagged data has been associated with the tagged data closely, and further build traction matrix to clear the right value. In the experiment, set gravity factor λ=1.2 , the specific steps are as follows:

Input: The sample data set $X=\{x_i \mid 1\le i\le n\}$, and $x_i \in S^Q$, define k clusters randomly, tagged sequence sets P , untagged sequence sets U , $P \cup U$=X .

Output: The clustering results are the objective function of the minimum k divided set: $\{X_m \mid_{m=1}^{k}\}$.

Step 1: Initialize the sample set closure Center;

Step 2: Using the initialized closure center to calculate the cluster center;

Step 3: Formula(2) also use closure replacement, namely using m_{ij} to represent $TF(\bar{x}_i, \bar{x}_j)$, and based

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

on it to establish traction matrix M in between the data points:

$$M = \begin{bmatrix} m_{11} & m_{12} & \cdots & m_{1k} \\ m_{21} & m_{22} & \cdots & m_{2k} \\ \vdots & & & \\ m_{n1} & m_{n2} & \cdots & m_{nk} \end{bmatrix} \quad (3)$$

In the formula, m_{ij} represent the traction force of $\overline{x_j}$ to $\overline{x_i}$, and satisfy formula (3).

Step 4: In the process of clustering, if $\overline{x_i} \in U$, and $\overline{x_i} \in C_i$, then $\overline{x_i}$ belong to the class i; else if $\overline{x_i} \in U$, and $\overline{x_i} \notin C_i$, then $\overline{x_i}$ belong to the Cluster C_k, and the limiting condition of the cluster to be defined:

$$C_k = \{(1-\delta)D_{x_i}^{\ 2} + \frac{\sum_{j=1}^{k} |\overline{x_i} - Z_j|^2}{m_{ik}k}\} \quad (4)$$

In the formula, δ is a correction factor, $\in (0,1)$.

Step 5: According to the formula (4) recalculate the cluster center, $Z_m^{i+1} = \frac{\sum \overline{x}}{X_m^{i+1}}, m = 1, 2, \cdots, k$.

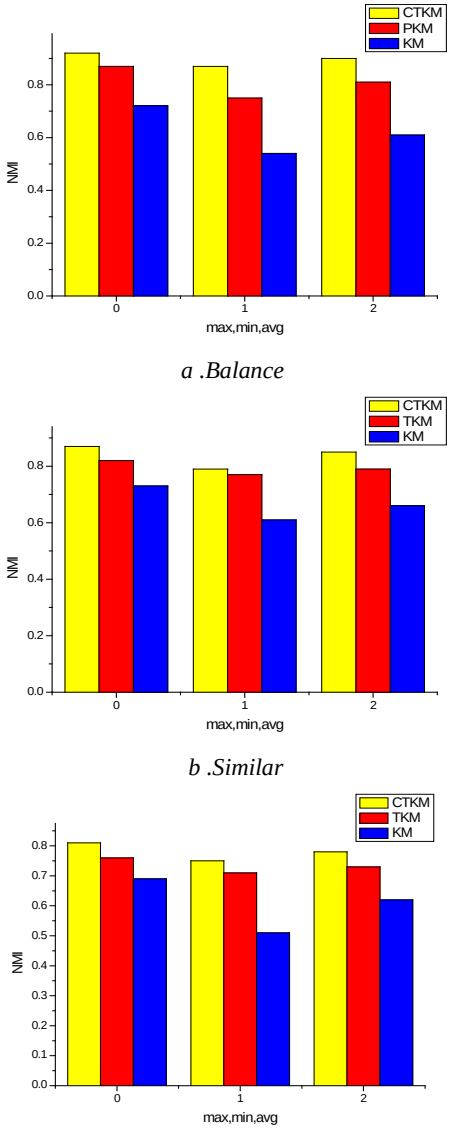
Step 6: If the condition $Z_m^{i+1} = Z_m^i$ is satisfied, the algorithm ends; else update iteration factor, to make $l = l + 1$, the process constantly back and forth until convergence.

The initial number of clusters is set in the K-means algorithm, and the cluster centers are selected randomly, therefore, the choice of the initial value will greatly influence the results of clustering. The introduction of the closure replacement in the text makes the cluster center under the supervision obtained greater credibility, while reducing the complexity of the algorithm; the introduction of traction in between the data points can also be effective to prevent the deviation of the individual untagged data clustering.

4. EXPERIMENTAL ANALYSIS

This paper involved three kinds of algorithm: the traditional K-means algorithm, denoted as KM; introducing the traction force of the K-means algorithm, denoted as TKM, as well as proposed in this paper CTKM algorithm based on the closure replacement. The experimental data derived from the three data sets in the universal database UCI: Balance, Similar and Simple. Use two quantitative indexes: NMI (normalized mutual information) and the clustering accuracy. 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.

The parameter settings as follows: $k = 3, \delta = 0.5$, each algorithm is run 15 times in the data set, figure 3 is a schematic diagram of the effect, respectively, showing the calculated maximum, minimum and mean:



a .Balance

b .Similar

c .Simple

Figure3: The Comparison Chart Of NMI Indicators

From the comparison in Figure 3, it is not difficult to find that introducing the closure replacement for the K-means algorithm clustering effect is significantly stronger than the other two

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

algorithms. In the case of the KM, The algorithm TKM and CTKM of results relatively close, the major difference of the two is the closure replacement, using the closure center instead of the ordinary sample point, the algorithm complexity is reduced, not only the time efficiency is improved, and the traction force between the sample points reflects obvious. With the lowest value of the three analysis, NMI value of the improved algorithm has a great improvement in the level, illustrate the isolated point classification higher accuracy, the clustering result of the algorithm is better.

In order to eliminate the impact of the supervision information randomness, according to the five groups different sets of data , three kinds of algorithm each run 5 times to calculate the accuracy of clustering, and each run randomly generated the tagged data, a number of categories are defined as 3 , the experimental results as shown in figure 4:

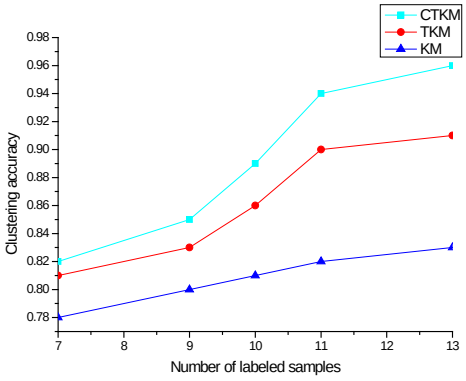


Figure 4: Clustering Accuracy

It can be seen by the experimental results of Figure 4: the improved algorithm of TKM and CTKM accuracy rate than KM increased significantly; taking into consideration the influence of the traction force between the data points, when the supervision information is small, the difference was not very obvious; With the increase in the supervision data, the gap of the accuracy rate between the algorithm is gradually widened, therefore, the influence of the tagged data to the untagged data has increased with the increase in the number. In addition, due to the closure replacement thinking adopted in the algorithm CTKM, isolated point can be accurately classified, the cluster effect is better.

5. CONCLUSION

In this paper, based on the thinking of semi-supervised clustering, we presents the closure replacement improved K-means algorithm to solve the irregular cluster clustering. The algorithm takes advantage of a small amount of the supervised information of tagged data, introduces the traction impact factor between the data points, at the same time, with the center of the closure instead of the original sample points, to simplify the process of clustering, also to eliminate the limitations of the K-means on the irregular cluster clustering. Experimental results indicate that the improved algorithm is much more efficient and simple than other algorithms on the comparison of the two indicators of the NMI and clustering accuracy.

REFRENCES:

[1] Chin-Chun Chang, Hsin-Yi Chen, "Semi-supervised clustering with discriminative random fields", *Pattern Recognition*, 2012, 45(12): 4402-4413.

[2] DoganayMC,Pedersen TB, SayginY, SavaşE, LeviA, "Distributed privacy preserving k-means clustering with additive secret sharing", *PAIS '08 Proceedings of the 2008 international workshop on Privacy and anonymity in information society*, 2008,3-11.

[3] Yeming Hu,Evangelos E. Milios, James Blustein, "Enhancing semi-supervised document clustering with feature supervision", *SAC '12 Proceedings of the 27th Annual ACM Symposium on Applied Computing*, 2012, 929-936.

[4] Guobiao Hu, Shuigeng Zhou, Jihong Guan, Xiaohua Hu, "Towards effective document clustering: A constrained K-means based approach", *Information Processing and Managemen*t, 2008, 44(4):1397-1409.

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

[6] Vid Podpečan, Miha Grčar,Nada Lavrač, "Semi-supervised constrained clustering: an expert-guided data analysis methodology", *PRICAI'10 Proceedings of the 11th Pacific Rim international conference on Trends in artificial intelligence*, 2010:219-230.

[7] Faußer S, Schwenker F, "Semi-Supervised kernel clustering with sample-to-cluster weights", *PSL'11 Proceedings of the First IAPR TC3 conference on Partially Supervised Learning*, 2011, 72-81.

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

[8] Tianhao Wu, Pottenger WM, "A semi-supervised active learning algorithm for information extraction from textual data: Research Articles", *ALNLP '10 Proceedings of the NAACL HLT 2010 Workshop on Active Learning for Natural Language Processing*, 2010, 10-17.

[9] ChristianBöhm, Claudia Plant. HISSCLU: a hierarchical density-based method for semi-supervised clustering", *EDBT '08 Proceedings of the 11th international conference on Extending database technology: Advances in database technology*, 2008: 440-451.

[10] ConstantinopoulosC, LikasA, "Semi-supervised and active learning with the probabilistic RBF classifier", *Neurocomputing*, 2008, 13-15(71): 2489-2498.

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