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 13 (11): 1786-1792, 2014
ISSN 1812-5638 / DOI: 10.3923/itj.2014.1786.1792

Research of Clustering Evaluation Index for User Grouping

[1]Jin Du, [2]Jun Hao and [1]Xiangmo Zhao
[1]School of Information Engineering, Chang'an University, Xi'an, 710064, People Republic's of China
[2]Shaanxi Xitie Real Estate Development Co. Ltd., Xi'an, 710064, People Republic's of China

Abstract: Some research had devoted to users grouping based on the clustering method. However, in many cases, the distribution of user's character is unknown, so that it is difficult to decide which algorithm is more suitable for user grouping. In this study an improved clustering evaluation index named as SSDS (inner-cluster Scattering, extra-cluster Separation, Density of centroids, balance of Size), was proposed. SSDS is able to select suitable algorithm for a data set and obtain the optimal clustering scheme though balancing among the weight of density, the degree of scattering and the size of clusters. Results of experiment for learner grouping indicated that this evaluation index method is feasible and efficient. In this experiment, the FarthestFirst algorithm with number of clusters set as 5 is the best clustering scheme.

Key words: Data mining, clustering evaluation index, user grouping, personalized service

INTRODUCTION

Nowadays, the personalized service pushing has become a promising field for web-based application such as e-learning, e-business, intelligent searching engine and so on. In Internet, the abundant resources are available and the research shows that the web-based application pushing personalized service according to the user's background, interesting, character, or custom, will improve service quality and gain more economic benefit (Zou and Ren, 2012).

A pivotal issue in personalized information system is that how the users were divided into some groups properly according to their characteristic in some views. For example, it is known that providing the personalized learning content and strategy will improve learners' mark efficiently. However, it is too difficult to be implemented that building a personalized information system which can meet the various type of personality and every single individual. The general method is to design a system for each user groups as a whole, within which user have similar characteristics. Thus, most of users would enjoy appropriate personalized service. On the other hand, the complexity of the building personalized information system would be reduced.

Generally, classifying and clustering, two main popular techniques of data mining, has been utilized for users grouping. The former trains the classifier by samples which has been labeled by experts beforehand and the classifier then divides new users according to their personality (Li and Ming, 2012). However, if the corresponding domain theory about users grouping is absent, so that the training data set with right label is unavailable. Comparably, the clustering method should be adopted. Clustering can assembles the similar objects into the same group according to the similarity distance between different objects, so the clustering result could describe the real distribution of data set.

However, clustering is an unsupervised analysis technique, which can always divide data set into several clusters no matter what algorithm adopted and what parameter set. Actually, different clustering algorithm will generates different clusters. Which algorithm is suitable for the requirement of application and which scheme can describe the real distribute of this data set? These questions must be answered above all. As the distribution of data set is unknown, to deal with these problems only by means of clustering validity assessment index, which could justify whether the clusters partition coincide with the real distribution of data set.

In this study, a novel clustering validity assessment method: SSDS was proposed and applied in e-learner grouping. Through this method, the suitable clustering algorithm and clustering scheme was found. Furthermore, the potential important characters of learner group, which would be taken as the basis for personalized learning service, can be concluded:

- Proposed an improved data oriented clustering validity assessment index: SSDS, which has the ability to select most suitable algorithm among several given clustering algorithms according to the distribution of data set
- Demonstrated the effectiveness of SSDS in experiment

Corresponding Author: Jin Du, School of Information Engineering, Chang'an University, Xi'an,710064, People Republic's of China

THE CURRENT EVALUATION INDICES FOR CLUSTERING VALIDITY

Background: Research shows that three kinds of clustering evaluation index are used: external measurement, internal measurement and relative measurement. Based on index's physical speciality, the optimal number of clusters often found at max/min value by assessing index. The performance of assessment is influenced by several factors, such as the data distribution, potential structure, features of data and algorithm selection (Jin *et al*., 2006). In addition, the features of data set would largely affect the performance of the algorithm, therefore the invalid algorithm will restrict the capability of assessment.

Weiguo Sheng presented that single estimate parameter can not keep its advantage on performance for all data set, so he proposes a method which sum up several popular assessing index (include 1/DB, SIL, VD, V33, CH and PBM) with weight to generate a new combined index (Shen *et al*., 2005). Aim at high-dimensional data set, Minho Kim proposed a method to release curse of high dimension (Kim *et al*., 2004), by replacing the Euclid distance with Manhattan segmental distance. The experiment shows that the performance of estimation was upgraded after modified. In his research, the purity was put forward to estimate the accuracy of data separating. Another work of Kim is to traditional clustering assessment modification in design principle. They replace the average of sum with max value to modify V_u and V_{SV}. However, in their experiment, they just examined the validity of one algorithm, which can not indicate the ability to select the right clustering algorithm so that it may avoid the phenomenon "right number of clusters but wrong partition".

S_Dbw*: S_Dbw* (Kim and Lee, 2003) is a novel clustering assessing index proposed by Youngok Kim and Soowon Lee. S_Dbw* improved the S_Dbw in the calculation of inner distance of cluster and outer distance between different clusters.

In data set S, giving a random point m, the density of m is defined as:

$$density^*(m) = \sum_{i=1}^{n} f^*(x_i, m) \tag{1}$$

where, n is the number of samples in s and:

$$f^*(x,m) = \begin{cases} 1 : CI^p \le d(x^p, m^p) \le CI^p, (\forall p, 1 \le p \le k) \\ 0 : otherwise \end{cases} \tag{2}$$

where, k is the number of dimensions of data set. CI^p is the confidence interval of the pth dimension. When confidence interval is set as 0.95, the CI^p is defined as:

$$CI^p = u^p \pm (1.96 \times \frac{\sigma^p}{\sqrt{n}}) \tag{3}$$

where, u^p up and σ^p is the mean and variance of the pth dimension.

Let m_{ij} denote the center of segment between clusters centroids C_i and C_j, the items in the formula 4 are replaced with follows:

$$u_{ij}^p = \frac{u_i^p + u_j^p}{2}, \sigma_{ij}^p = \frac{\sigma_i^p + \sigma_j^p}{2}, n_{ij} = n_i + n_j \tag{4}$$

Hence, the inner density of cluster Dens_bw* (c) is defined as:

- The scattering degree of clusters Scat* (c) is calculated by following formula:

$$Dens\_bw^*(c) = \frac{1}{c(c-1)} \sum_{i=1}^{c} \left[\sum_{\substack{j=1 \\ i \ne j}}^{c} \frac{density^*(m_{ij})}{\max\{density^*(V_i), density^*(V_j)\}} \right] \tag{5}$$

$$Scat^*(c) = \frac{1}{c} \sum_{i=1}^{c} \frac{n - ni}{n} \left(\left\| \sigma(V_i)^2 \right\| / \left\| \sigma(S)^2 \right\| \right) \tag{6}$$

where, $\|x\| = (xx^T)^{1/2}$

At last:

$$S\_Dbw^*(c) = Dens\_bw^*(c) + Scat^*(c) \tag{7}$$

The minimum value of S_Dbw* would indicates the optimal clustering result.

The comparison between S_Dbw* and S_Dbw shows that former have the better performance than latter one. Some research points out that the performance of S_Dbw is better than RS, RMSSTD, DB and SD, thereinto, the DB was regarded as a robust assessment index (Shim *et al*., 2005). So it is deduced that S_Dbw* is an excellent clustering assessing index.

AN IMPROVED EVALUATION INDEX

The primary function of clustering is to recognize clusters, which is dense data section with high dimension. Therefore, plenty of clustering algorithms were proposed to fulfill the demand of different domain and different data set with special attributes (e.g., high-dimension or random distribution). Generally, the shape of data set is unknown and little domain knowledge can be used for reference. So, how to choose an algorithm will be the chief question

above all. Based on S_Dbw*, an improved assessing index called SSDS is proposed in this study. SSDS consider the scattering, separation, size and density of clusters all together and is enable to select clustering algorithm according to the character of data set.

Definition of SSDS: For data set S, a suitable algorithm should be selected from a set of available clustering algorithms. The meaning of "suitable" include: (1) Obtain the accurate number of clusters, (2) Generate perfect partition of data set S, this partition must coincide with the real distribution as good as possible.

SSDS take the density as the essential standard to judge whether the algorithm is appropriate for data set. Here, the Dens (c) was defined as:

$$Dens(c)=\begin{cases}\frac{1}{c}\sum_{i=1}^{c} n_i, & \text{if } density^*(Vi)-\frac{1}{c-1}\sum_{\substack{j=1\\ j\neq i}}^{c} density^*(m_{ij})\le 0\\ \frac{1}{c}\sum_{i=1}^{c}\frac{n_i}{density^*(V_i)-\frac{1}{c-1}\sum_{\substack{j=1\\ j\neq i}}^{c} density^*(m_{ij})}, & \text{otherwise}\end{cases} \quad (8)$$

where, density* (V_i) is the density of centroids in the ith cluster, density* (m_{ij}) is the density of center of segment between the ith cluster and the jth one and noted as d* (V_i) and d* (m_{ij}). The minimum of Dens (c) indicate that optimal clustering algorithm and pattern are generated.

As is known, if the size of clusters is widely different, the result of clustering will be invalid in real application. Here, the size of c was defined as:

$$Size(c)=\frac{\sum_{i=1}^{c}\left|n_i-\bar{n}\right|}{n},\ \bar{n}=n/c \quad (9)$$

where, n_i is the number of samples in the ith cluster.

With the concept of S_Dbw*, in formula (3), S_Dbw* using the standard normal distribution range parameter evaluation with $\alpha = 0.05$ but α is fixed for all data set. In this works, a variable marked as α, which was the parameter for confidence interval and utilized to measure the compacting and scattering degree of clusters structure adaptively was introduced. Here, α can be adjusted according to the character of data set. The S_Dbw* is improved as S_Dbw* (c,a) and the formula 3 is modified as:

$$CI^p(\alpha)=u^p\pm(Z\times\frac{\sigma^p}{\sqrt{n}}) \quad (10)$$

where, Z is the coefficient for α in standard normal distribution. The S_Dbw* after modified would be calculated by following formula:

Table 1: Table (Z, α), the corresponding values of Z to α in standard normal distribution

α	Z_α
0.8	0.26
0.5	0.68
0.3	1.04
0.1	1.65

α: Value of confidence interval, Z: Coefficient to α in standard normal distribution

$$S\_Dbw^*(c,\alpha)=Dens\_bw^*(c,\alpha)+Scat^*(c) \quad (11)$$

The values of Z corresponding to α were shown as Table 1.

The points taking part in the calculating for density of m in formula (1), are named as neighbors of m and the interval which are bounded by CIp for k dimensions of m, were named as neighbor interval. Variable α decide the size of m neighbor interval in assessment. For data set S, the first step is to adjust α by SSDS, Once α is confirmed, it don't have to be changed anymore. Especially, if d* (Vi) worked out by all the algorithms equals n (n is the size of S), the α should be increased to constrict the confidence interval. If d* (Vi) equal 0, α should be decreased to enlarge the confidence interval.

Finally, according to the Eq. 8, 9 and 11, the index of SSDS can be defined as:

$$SSDS(c)=\beta.Dens(c)+\gamma.Size(c)+S\_Dbw^*(c,\alpha) \quad (12)$$

where, both β and γ is balance factor, which can adjust the importance of different estimating aspects in clustering assessment. Generally, to avoid some part of indexes dominating SSDS, the magnitude of two parts ahead are adjusted according to that of S_Dbw* (c,a). Similar withα, β and γ are adjusted in the beginning of assessment of a particular data set and change free till data set is changed.

The illustration for validity of SSDS in algorithm selection: In this subsection, it will be proved that the function of component Dens (c) make the SSDS have the capability to select an optimal algorithm. As Fig. 1a-d illustrated, here are an example data set S which was clustered by a set of algorithms labeled as {$Algo_i$} ($1\le i\le 4$), those clustering consequences were various with different algorithm. The Fig. 1a can be regarded as an actual partition for S by $Algo_1$. Figure 1b-d describe various clustering partitions by $Algo_{2\text{-}4}$. From Fig. 1a-d, the centroid in cluster and center between two different clusters were marked by dots. Here, supposing the all of data dots distributed in all zones symmetrically.

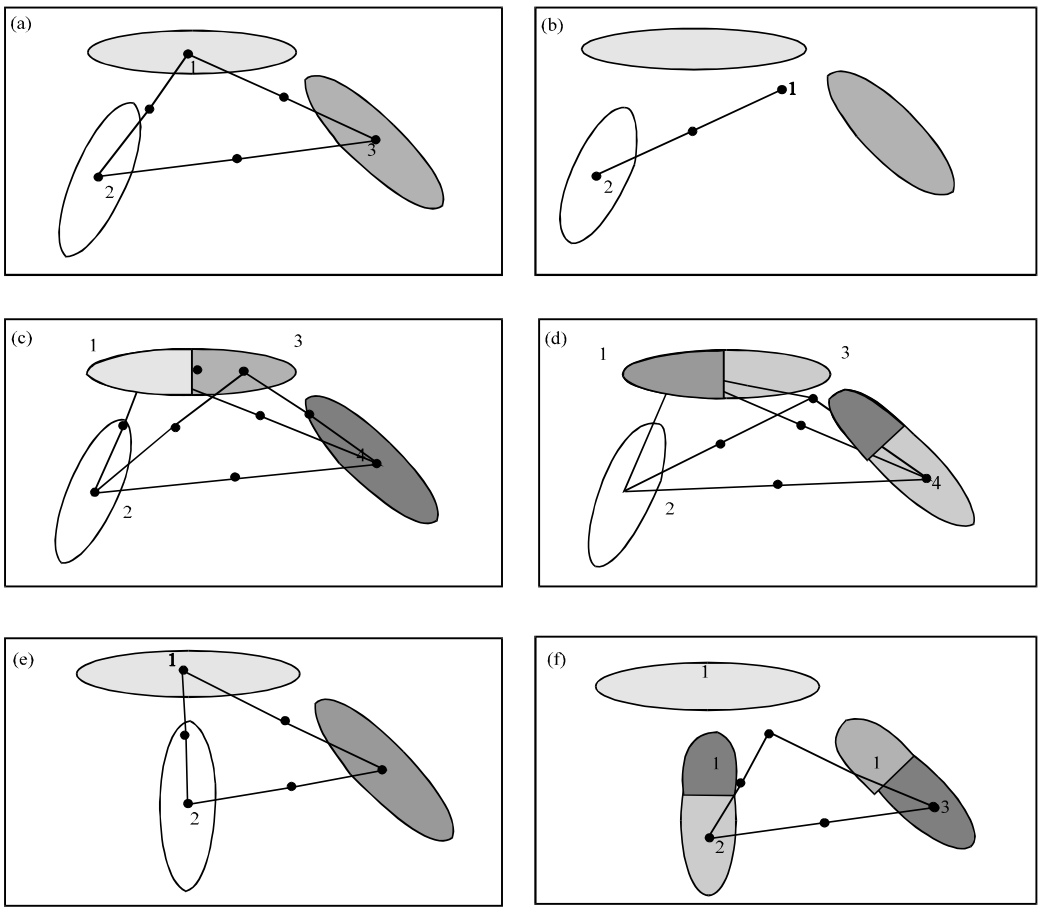

Fig. 1(a-f): Various clustering result for data set S with different algorithm, (a) Actual partition, (b) Partition over combined, (c) Partition1 over split, (d) Partition 2 over split, (e) Correct cluster number and correct partition and (f) Correct cluster number but wrong partition

Firstly, the partition generated by $Algo_2$ has too many data combination. As Fig. 1b showed that, n_1 is very great but the data dots distributed among α-neighbor field between V_1 and m_{12} is few, so the value of $d^*(V_1)$ and $d^*(m_{12})$ is very little even close to zero. According to Eq. 9:

$$\frac{n_1}{d^*(V_1) - d^*(m_{12})} \approx n_1$$

Dens (c) and SSDS will become very large, which indicate that algorithm $Algo_2$ is not suitable for S.

Secondly, as Fig. 1c showed, $Algo_3$ split data set S so much that 4 clusters were generated. As a result, there are lot of dots in the α-neighbor field of middle dot m_{13} between V_1 and V_2, so $d^*(m_{13})$ is close to or equals $d^*(V_1)$ and $d^*(V_3)$. Here:

$$d^*(V_1) - \sum_{j=2}^{4} d^*(m_{1j}) \approx 0$$

this phenomena was called as "middle density is first", which will result in Dens (c) and SSDS become very large, so is neither the algorithm $Algo_3$ suitable for S.

As same as $Algo_3$, the number of clusters divided with $Algo_4$ is too more and the phenomena "density of cluster center has priority" still existed in $cluster_3$. In this situation, the $d^*(V_3)$ is small but $d^*(m_{13})$ and $d^*(m_{34})$ are very large so that Dens (c) and SSDS become very large, therefore the $Algo_4$ still was denied as Fig. 1d showed.

Only the data set S was partitioned as Fig. 1a showed, d* (V) is large and d* (m) is little, therefor the values of Dens (c) and SSDS were decreased obviously.

This phenomena was called as "centroid density is first", which indicated $Algo_1$ is most suitable for S.

In some case that the clusters number is right but the partition is wrong, SSDS also can judge which algorithm is adapt to data set accurately. Figure 1e-f illustrated two different results generated by $Algo_1$ and $Algo_2$ for data set S, both them include three clusters, meanwhile all centroids within clusters and centroids between clusters were marked by dot. Here, the first result, as Fig. 1e showed, is the correct partition but second result as Fig. 1f showed may be wrong. Obviously, the $cluster_1$ in Fig. 1f is "middle density is first", so SSDS denied $Algo_2$ and selected $Algo_1$ as the optimal algorithm for data set S.

Therefore, aiming to data set S, SSDS can find out the optimal clustering algorithm which would generate correct numbers of clusters and achieve best partition.

The progress of algorithm selecting by SSDS: The capability to auto selecting clustering algorithm by SSDS was proved, here, the progress of algorithm selecting according to data set S by SSDS were described as following:

Algorithm: Procedure SSDS

Input: data set S, all of cluttering results { $Algo_i$ $(p_1,p_2,\ldots,p_k)$}, Table (α, Z)

Output: optimal algorithm Algo*, optimal input paramters of Algo* (p_1*,p_2*,…,p_k*)

//select a probe algorithm randomly and use its Dens(c) and Size(c) to adjust α,β,γ

Select an Algo$\in${$Algo_i$}

$\alpha \leftarrow 0.05$;

while (max{Denssity(Vi)} < n* 20%){
    decrease α according to Table(Z, α) ;
    if Denssity(Vi) don't change any more break;
}

while (max {Denssity(Vi)} > n* 80%){
    increase α according to Table(α, Z);
}

$\beta \leftarrow \frac{1}{\text{magnitude (Dens/S\_Dbw*)}}$; //magnitude () get the order of magnitude

$\gamma \leftarrow \frac{1}{10 * \text{magnitude (Sizes/S\_Dbw*)}}$;//if without special require for balance of cluster size

End

//calculate values of SSDS for different algorithm and different parameters.

For Algo, $(p_1,p_2,\ldots,p_k)\in$ { $Algo_i$ $(p_1,p_2,\ldots,p_k)$}
    calculate SSDS ($Algo_i$, $(p_1,p_2,\ldots,p_k)$);

End for

//the least value of SSDS indicate that the optimal algorithm is selected and the best partition wasgenerated.

For SSDS ($Algo_i$, $(p_1,p_2,\ldots,p_k)$)
  SSDS* ←min{ SSDS ($Algo_i$, $(p_1,p_2,\ldots,p_k)$)};
    Algo* ←SSDS*.Algo;
    $(p_1$*,p_2*,…,p_k*)←SSDS*.$(p_1,p_2,\ldots,p_k)$;

End for

Return Algo*,(p_1*,p_2*,…,p_k*);

Although, SSDS have two more calculating works for Dens (c) and Size (c) than S_Dbw*, the middle consequence of can be used in calculating progress entirely, so both two indexes have similar time cost.

EXPERIMENTS AND DISCUSSION

In this study, the experiment of personality clustering was discussed. The personality were represented by Cattle 16PF (Personality Factor) (Cattell and Schuerger, 2003), which is famous psychics theory and describes person's character with 16 different personality attributes. Therefore, the user Personality Model (PM) can be represented as a vector with 16 dimensions such as:

$$\overrightarrow{PM} = < A,B,C,E,G,H,I,L,M,N,O,Ql,Q2,Q3,Q4 >$$

Here, the value of each dimension is from 1-10.

After data cleaning, 256 students from Chang'an University whose personality data were collected and data set S was built. However, the shape of personality data distribution is unknown. Here, a data mining tool, WEKA, was used to facilitate the experiment. But a few further developments were attached to Weka for the present work.

In order to get the most suitable algorithm and accurate clustering scheme, three popular algorithms: K-means (K), EM and FarthestFirst (FF) were adopted in experiments. Set number of clusters from 4-7 and begin seed from 20 points which distributed evenly. The best result of every algorithm was shown as Table 2. Here, the d (V) and d (m) were replaced by $\bar{d}(V_i)$ and $\bar{d}(m_{ij})$.

In column 3, the Min (SSDS) of FarthestFirst have the smallest value, which is indicated that the result 1 (row1) is the optimal clustering scheme and FF is the most suitable algorithm for personality data set. The $\bar{d}(V_i)$ of FF is greater than that of other algorithm's and follows the principle "centroid density is first". The experiment showed the result 1 and 2 are very similar on Min (SSDS)

Table 2: Comparison among farthest first, EM and K-means, whose clustering results were evaluated by SSDS ($\alpha = 0.05$, $\beta = 0.01$, $\gamma = 0.1$)

Clustering algorithm	Min (SSDS)	Clusters	Size of clusters	$\bar{d}(V_i)$	$\bar{d}(m_{ij})$
Farthest first	0.2435	5	55, 32, 55, 107, 16	73.4	31.8
Farthest first	0.2485	4	64, 34, 55, 112	81.5	38.5
EM	0.3273	5	87, 103, 34, 34, 7	14.6	1.2
K-means	0.3596	7	46, 40, 44, 41, 33, 43, 18	10.4	3.2

α: Value of confidence interval; β: Balance factor for density; γ: Balance factor for size

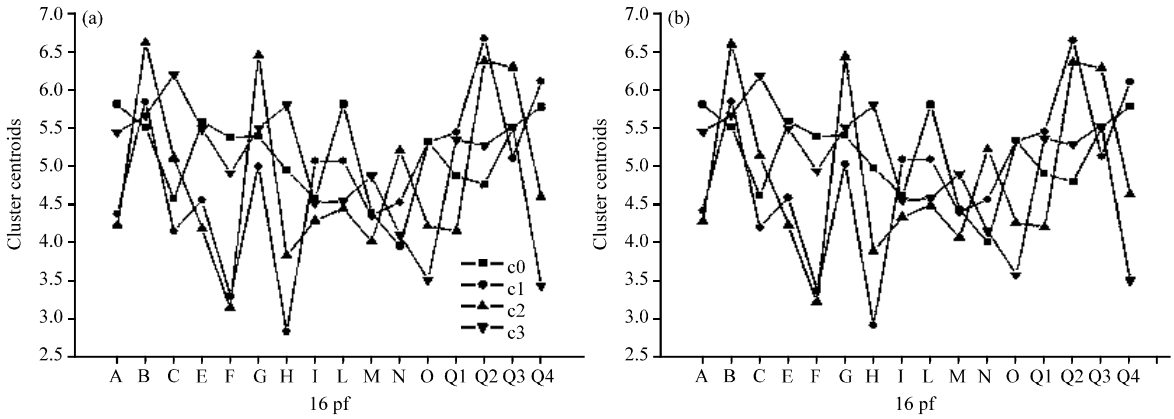

Fig. 2(a-b): Distribution of cluster centroids on each attributes of 16PF (a) Five clusters and (b) Four clusters

(only different in 0.005), it accords with the previous study that when the examples were grouped into 4 or 5 clusters and the optimal clustering pattern is obtained.

Figure 2a-b revealed the distribution of clusters centroids on each attributes of personality model (16PF) in result 1 and result 2.

As a whole, $\overline{d}(V_i)$ and $\overline{d}(m_{ij})$ is very small in result 3 and result 4, their value seldom increased when α was decreased to 0.0125 to expand confident interval. The reason may be that personality data is so loose or unbalanced that EM and K-means can not generate good result for this kind of data set. The comparison among three algorithms shows that algorithm selection is of great importance for a particular data set.

CONCLUSION

After comparing some classic clustering evaluation indexes, a novel assessment index: SSDS was proposed. SSDS integrates the advantages of several essential indexes (including inner-cluster Scattering, extra-cluster Separation, Density of centroids /middle points, balance of Size). The research indicated that SSDS is a data-oriented index which can select an optimal algorithm by estimating the precision of candidate algorithms. In application on learner grouping based on personality clustering, according to SSDS, it is demonstrated that FarthestFirst had excellent performance and that the five clusters represents divided will represent the real distribution of personality of learners accurately. The experiment shows that, aiming at the some lack of theory to support user grouping based on their features, a clustering method with validity assessment will be an effective measure.

ACKNOWLEDGMENTS

This research is partially supported by the Nature Science Foundation of Shaanxi Province (2009JQ8002), The special Fund for Basic Scientific Research of Central Colleges and the Special Fund of Basic Research Support Program of Chang'an University (CHD2011JC021), Shaanxi Engineering and Technical Research Center for Road and Traffic Intelligent Detection.

REFERENCES

Cattell, H.E.P. and J.M. Schuerger, 2003. Essentials of 16PF Assessment. John Wiley and Sons Inc., New York, USA., ISBN-13: 9780471474135, Pages: 320.

Jin, D., Q. Zheng, D. Jiao and G. Zhiyong, 2006. A method for learner grouping based on personality clustering. Proceedings of the 10th International Conference on Computer Supported Cooperative Work in Design, May 3-5, 2006, Nanjing, China, pp: 1-6.

Kim, Y. and S. Lee, 2003. A clustering validity assessment index. Proceedings of the 7th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining, April 30-May 2, 2003, Seoul, Korea, pp: 602-608.

Kim, K., H. Yoo and R.S. Ramakrishna, 2004. Cluster validation for high-dimensional datasets. Proceedings of the 11th International Conference on Artificial Intelligence: Methodology, Systems and Applications, September 2-4, 2004, Varna, Bulgaria, pp: 178-187.

Li, J. and D. Ming, 2012. Research of the users personalized model in Web mining. Proceedings of the International Conference on Computer Science and Service System, July 27-29, 2012, Nanjing, China, pp: 1572-1574.

Shen, W.G., S. Swift, L. Zhang and X. Liu, 2005. A weighted sum validity function for clustering with a hybrid niching genetic algorithm. IEEE Trans. Syst. Man Cybern. Part B: Cybern., 35: 1156-1167.

Shim, Y.S., J. Chung and I.C. Choi, 2005. A comparison study of cluster validity indices using a nonhierarchical clustering algorithm. Proceedings of the International Conference on Computational Intelligence for Modelling, Control and Automation, November 28-30, 2005, Vienna, Austria, pp: 199-204.

Zou, L.W. and G.W. Ren, 2012. The data mining algorithm analysis for personalized service. Proceedings of the 4th International Conference on Multimedia Information Networking and Security, November 2-4, 2012, Nanjing, China, pp: 332-335.