Automatic K Selection Method for the K - means Algorithm

Xiuli Shao
College of Computer and Control Engineering.
Nankai University
Tianjin, China

Huichao Lee
College of Computer and Control Engineering.
Nankai University
Tianjin, China

Yiwei Liu
The Fu Foundation School of Engineering and Applied Science
Columbia University
New York, NY, USA

Bo Shen
Tianjin Port Information Technology co. ltd
Tianjin, China

Abstract—**The key problem of the K-means clustering algorithm is to ascertain the appropriate number of clusters, that is K value, usually by ceaseless experiments to obtain it, the voting mechanism method was introduced in this paper to automatically determine the optimal number of clusters in a dataset. The voting mechanism method was introduced in this paper to automatically determine the optimal number of clusters in a dataset. In this method, the symbol K is used to represent the number of clusters, and a series candidate sets of K were obtained by using different validity indexes. Then, the frequency of each K was calculated, and we chose the highest frequency as the final number of clusters, that is the value of K. Using the method proposed in this paper to divide the real dataset of the ship tonnage, the experimental result is basically consistent with the rules of the actual operation, so the experimental verify the reliability and effectiveness of the approach presented in this paper.**

Keywords—ship tonnage clustering; K-means; determination of the number of clusters; voting mechanism; rule extraction

I. INTRODUCTION

Port enterprises as a part of the logistics, mainly provide the ship's shore docking and handling, cargo storage and storage services[1], have accumulated a large number of data failed to be fully utilized in the years of operation. With the advent of the big-data age and the development of business intelligence, port business managers hope to use the data mining model to classify the tonnage of incoming ships, so that the port enterprises can facilitate timely and reasonable arrangements, which is of great importance.to shorten the ship's time and cost in the port.

Through the preliminary research, It is found that the port business data is multi-source heterogeneous, and unlabeled. K-means is a typical clustering algorithm [2, 5], which aims to reveal a class structure in a dataset by partitioning it in an unsupervised manner. It is attractive in practice, because it is simple and generally fast. However, a natural question in cluster analysis is how many clusters are appropriate for the description of a given system [20, 21].

For clustering result, it is always evaluated by some validity indexes. A large number of validity indexes have been presented in the literature. Some indices are based on considering the compactness within each cluster or the separation between clusters [23–26], like the weighted distance [3, 4]. Some are based on the information entropy or based on measuring in-group proportion of objects [21, 27]. For example, H. Xiong, J. Wu and J. Chen [22] introduce the entropy measure, an external clustering validation measure, has the favorite on the clustering algorithms which tend to reduce high variation on the cluster sizes. P. J. Rousseeuw [15] and Etienne Lord [19] both use the average silhouettes width to interpret and analyze the clusters.

In the experiment of ship tonnage division, it was found when using average centroid distance to select K value, the weighted average centroid distance within a cluster and the weighted average centroid distance between clusters can not be optimized simultaneously. As the validity indexes to determine the number of clusters. The determination of the number of clusters requires manual tradeoff, which is laborious and reduces the practicality of the model. In order to solve this problem, we propose a selection model based on voting mechanism to determine the number of clusters automatically. Our inspiration of applying the voting mechanism comes from the idea of majority consensus used by Thomas [6]. The voting mechanism has been successfully applied to many problems, for instance, Gifford suggested using voting mechanisms for replicated data management [7], Daniel Barbara and Hector applied it in a faulty distributed system [8]. In this paper, evaluating the clustering results gained by the K-means algorithm from multiple angles by applying multiple validity indexes. What's more, this method was used to obtain the optimal classification of ships, and it was consistent with reality classification of ships. Therefore, it shows that the method can help us to get

Funding areas is Tianjin science and technology project：the key technology of big data (Grant No. 15ZXDSGX00090)

978-1-5386-1107-4/17/$31.00 ©2017 IEEE

the clustering results with the better compactness and relatively uniform sample distribution.

Section II utilizes K-means and two validation indexes (average centroid distance and silhouettes) on the ship data, analyzes the clustering results as well. The core of the paper is the third section, in which, a voting mechanism model for determining the number of clusters was introduced. In Section IV the experimental partition of the ship data was presented and its validity and reliability were discussed.

II. CLASSIFICATION OF SHIP TONNAGE BASED ON K-MEANS

Because the water depth of the port is different from the actual draft of the ship, In order to ensure the safety of ships，the port arranges the ships docked at the port to rush the loading and unloading promptly in the sea water high tide, and ships have to sail from the shore at ebb tide Because the port operations was usually scheduled based on the classify the tonnage of the vessel, the K-Means clustering algorithm based on voting mechanism method was used in this paper for classifying of inbound tonnage of ships, and carry out regular operation guidance for different vessels, facilitate the timely and scientific scheduling of ports, complete grabbing water operations, and ensure the orderly and safe operation of the port

A. Preprocessing of the basic watercraft data

In this paper, from the port dispatching command center, the historical data of the container ship's pre reporting during the period of 2015/12-/2016/12 is selected as the initial data of the ship clustering, totaling 4850.

The characteristics of each data sample, including the captain, the width of the boat, the fore draft and aft draft of the boat, the number of imported containers, the number of exported containers, the nature of the ship and the deadweight tons. According to the practical experience, there is obvious linear relationship between the captain and wide, before and after the draft of boat. These correlated features are redundant and can't provide large common information for data mining [10], and make out inaccurate distance calculation in the experiment. In practice, this discovery process should avoid redundancies with existing knowledge about class structures [9].So the original data needs to be dimensioned[17] and compressed, aiming at selecting the top non-redundant features such that the useful mutual information can be maximized.

In order to determine the correlation between the attributes in the data set, the Pearson correlation coefficient between the two features are calculated, and it turns out that the correlation coefficient between "ship length" and "ship width" is 0.95, "draft forward" and "draft after" is 0.93, "the number of imported containers" and "the number of export containers" is 0.7. The above results show that there is a strong correlation between the fields of the existing ship data.

After the PCA dimensionality reduction, the eight data attributes are reduced to 5 dimensions in the original data, and {F_1, F_2, F_3, F_4, F_5} are used to represent that 5 newly selected attributes in this paper.

B. Analysis of clustering results using different K

One main problem in applying the K-means clustering algorithm is selecting the appropriate number of clusters [13]. This problem is a very active field of research [11, 12, 14]. In this paper, we use the clustering performance evaluation indexes to select the appropriate K.

For the ship dataset $D=\{x_1,x_2,...,x_n\}$ given in Section A, when the value of K is fixed, the ships dataset after clustering is divided into $C=\{C_1,C_2,...,C_k\}$ The clustering center of C_t isx_t^c. The cluster weighted average centroid distance of a cluster C_t is calculated according to (1), and the weighted average centroid distance between clusters can be calculated according to (2).

$$\mathrm{avg\_dist}_{\mathrm{core\_in}}=\sum_{t=1}^{K} w_t \cdot \mathrm{avg}(C_t) \tag{1}$$

$$\mathrm{avg}(C_t)=\frac{1}{|C_t|}\sum_{1\le i\le |C_t|} \mathrm{dist}(x_i,x_t^c) \tag{2}$$

$$w_t=\frac{|C_t|}{\sum_{t=1}^{K}|C_t|} \tag{3}$$

where $avg(C_t)$ is the average of the distance between the samples in the cluster C_t and the cluster center, w_t is the weight value of the tth cluster C_t in the entire ship dataset D, that is, the proportion of the samples in the cluster C_t in the whole dataset.

$$\mathrm{avg\_dist}_{\mathrm{core\_out}}=\sum_{t=1}^{K}\frac{\sum_{c_j\in c,\, c_j\neq c_t}\mathrm{dist}(c_t,c_j)}{K-1}\cdot w_t \tag{4}$$

The trend of the weighted mean centroid distance in one cluster gained by using different K value is shown as Fig.1. By comparing the values of weighted mean centroid distances in the cluster corresponding to different K, it is found that the weighted average of the centroid distance in the cluster declines faster when k is 4, 5, 6, 7 respectively.

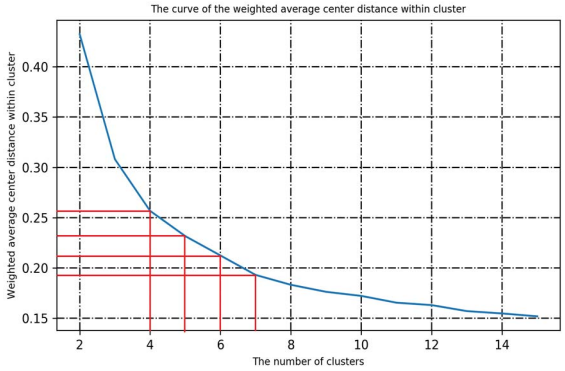

Fig. 1 The weighted average centroid distance of the cluster corresponding to the different K values

The variation curve of the weighted average centroid distance between clusters is as shown as Fig. 2. It can be seen from the figure that the weighted average centroid distance between clusters is the largest when k =5. Followed by k = 4, the average center of mass distance between clusters is lower than k = 5. When $k \geq 6$, the average center distance between clusters is relatively small.

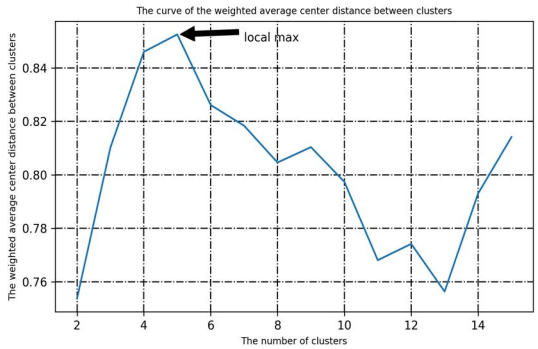

Fig. 2 The weighted average centroid distance between clusters corresponding to the different K

From Fig. 1 and Fig. 2, it can be found that k=7 is the more appropriate number of clusters if only the weight average centroid distance of one cluster was considered. It can obtain smaller cluster average centroid distance than other k value, which indicates that the distribution of samples in a cluster is more compact. However, the average centroid distance between different clusters become very small also when there are seven clusters, which indicates that the clusters are overlapped. When k = 5, the weighted average centroid distance between clusters is the largest, indicating that all clusters are relatively dispersed; but the corresponding average centroid distance of a cluster is also large, indicating that the cluster is not enough aggregation.

From the above analysis, the average centroid distance between clusters and the average centroid distance in one cluster can't be obtained optimally at same time. In order to obtain the optimal centroid distance, for selecting fit K value, it need to be artificially weighed, which is time-consuming and labor-intensive, and reduce the usefulness of the model.

C. Determination of K candidate set based on Silhouettes

A set of better k values obtained in Section B is taken as the candidate set $Set_1(Set_1=\{4,5,6,7\})$.In order to further determine the optimal number of clusters, In addition, the " silhouettes coefficient " is selected as the second validity evaluation index to evaluate the cluster partition as C={" "C" _"1" "," "C" _"2" ",…," "C" _"k" "}" of the ship dataset obtained in Section B.

The silhouettes coefficient $s(x_i)$ in (5) shows how well the ship sample x_i lie within its cluster [15]. If $s(x_i)$ is closed to +1, which means that x_i is appropriately clustered. If $s(x_i)$ is close to -1, which means that sample x_i is misclassified and maybe it would be more appropriate to clustered into its neighbor cluster. When $s(x_i)$ is close to 0 which means that x_i is on the border of two natural clusters, that is, the overlap between the clusters is serious. These can not be reliably assigned to the clustering results [16].

$$s(x_i)=\frac{b(x_i)-a(x_i)}{\max\{a(x_i),b(x_i)\}} \tag{5}$$

$a(x_i)$ in (5) denotes the average distance between ship sample x_i and other samples in the same cluster, and can be computed according to (6). $b(x_i)$ in (7) denotes the minimum value of the distances from the sample x_i to the other k-1 clusters, where $dist(x_i,C_k)$ can be calculated as (8).

$$a(x_i)=\frac{1}{|C_t|}\sum_{j=1}^{|C_t|} dist(x_i,x_j) \tag{6}$$

$$b(x_i)=\min\{dist(x_i,C_k)\,|\,x_i\in C_t,1\le k\le K,\ k\ne t\} \tag{7}$$

$$dist(x_i,C_k)=\frac{1}{|C_k|}\sum_{j=1}^{|C_k|} dist(x_{ti},x_{kj}),\ x_t\in C_k \tag{8}$$

$s(x_i)$ is only used to show how well a single object x_i matches the clustering at hand (that is, how well it has been classified).In order to determine the overall clustering effect for a certain K, the average silhouettes coefficient $avg(s(x_i))$ for all ship samples is computed according to (9).

$$avg(s(x_i))=\frac{\sum_i^n s(x_i)}{n} \tag{9}$$

The trend of the average of the silhouettes coefficient corresponding to different K is as shown as Fig. 3.

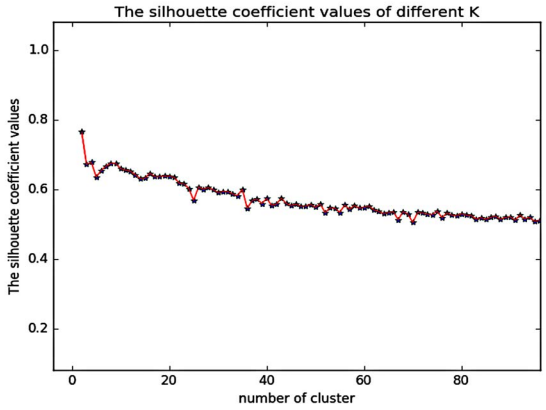

Fig.3 Variation curve of silhouettes score for different K

The horizontal axis represents the number of clusters, and the vertical axis represents the silhouettes coefficient value. With the number of clusters growing, the silhouettes coefficient value decreases, indicating that the result of clustering is getting worse. Therefore, in the experiments that uses the K-means method to classify ship tonnage, the number of clusters should not be too large.

The table I lists the top ten silhouettes score and the corresponding value of K. When the value of K is 2, 4 and 8 respectively, the corresponding silhouettes scores are ranked in the first three. Moreover, when the number of clusters is 5, the weighted mean centroid distance between clusters is optimal, but the silhouettes coefficient score is the lowest.

TABLE I RANKING TOP TEN SILHOUETTES AND CORRESPONDING K

Rank	K	$avg(s(x_i))$	Rank	K	$avg(s(x_i))$
1	2	0.765724	6	7	0.666205
2	4	0.677648	7	10	0.659301
3	8	0.675144	8	11	0.655128
4	9	0.674665	9	6	0.654429
5	3	0.671695	10	5	0.634971

The overall average silhouettes coefficient is not enough to fully determine the quality of the clustering results. The distribution of samples in each cluster is also important. The silhouettes of the different clusters are printed below each other. In this way the entire clustering can be displayed by means of a single plot, which enables us to distinguish 'clear-cut' clusters from 'weak' ones[15].

The distribution of samples in each cluster can be visually observed by the silhouette plot. The vertical axis represents the index of each cluster and the horizontal axis lists the values of the $s(x_i)$.

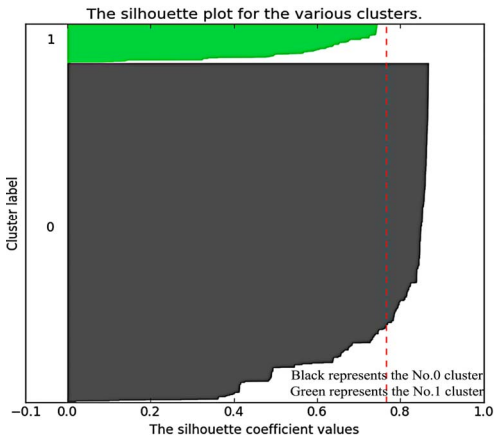

Fig.4 Silhouettes of a clustering with k=2 of the ship data

Figure 2 shows the silhouettes for the clustering into k = 2 clusters of the ship data. When K=2, the average silhouette coefficient is the highest, but the $s(x_i)$ value of the No.0 cluster is lower than the $\text{avg}(s(x_i))$ value of all clusters, indicating that the No.0 cluster contains too many samples, while No.1 cluster contains very few samples, that shows the cluster samples distribution are not compact enough.

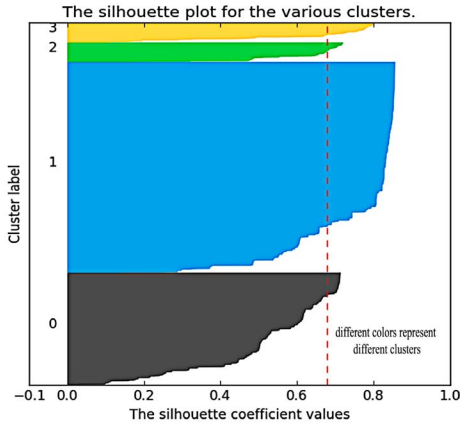

Fig. 5 Silhouettes of a clustering with k=4 of the ship data

Fig. 5 shows the silhouette plot of the same data, but now partitioned into four clusters (also by means of the K-means method). When K=4, there is no cluster with silhouettes less than average, so the effect of comprehensive clustering is better.

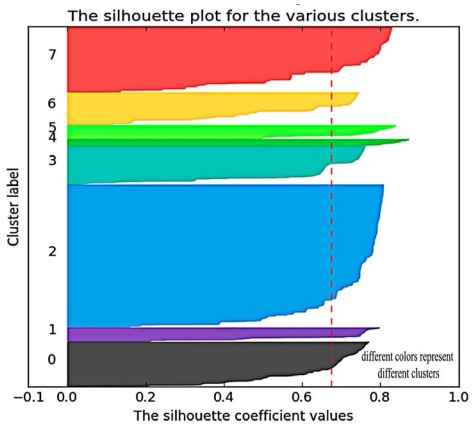

Fig. 6 Silhouettes of a clustering with k=8 of the ship data

When K = 8, the silhouette coefficient is 0.675144,which ranked third, and the variation of the silhouettes thickness of each cluster fluctuates greatly, and the number of samples among the clusters is not balanced.

Based on the above analysis, the candidate set Set_2={2, 4, 8} is obtained by using the "silhouettes" as the clustering performance evaluation index.

III. Determination of the Number of Clusters Method of K-Means Algorithm Based on Voting Mechanism

From the experimental results in Section II-B, it is clear that the average centroid distance in one cluster and the distance among clusters can not be optimized at the same time. The choice of the optimal number of clusters in K-means algorithm needs to be weighed artificially, which is not only time-consuming and labor-efficient but also reduces the usefulness of the model. Thus, in this paper, the selection model of the K value is established based on the voting mechanism in order to obtain the optimal K value automatically.

The model in this paper based on the principle of "minority obeying majority" in the voting mechanism, and the K with the largest frequency in all the candidate set is selected as the final number of clusters. The core idea of the model is to use multiple clustering validation indexes to evaluate the generated partition, to determine the optimal K and get the optimal partition of the ship data set. Compared with the traditional K-means clustering algorithm, the advantage of this model is that the final clustering number can be automatically determined without human intervention. To this end, the structure of the K selection model based on the voting mechanism is as shown in Fig. 7.

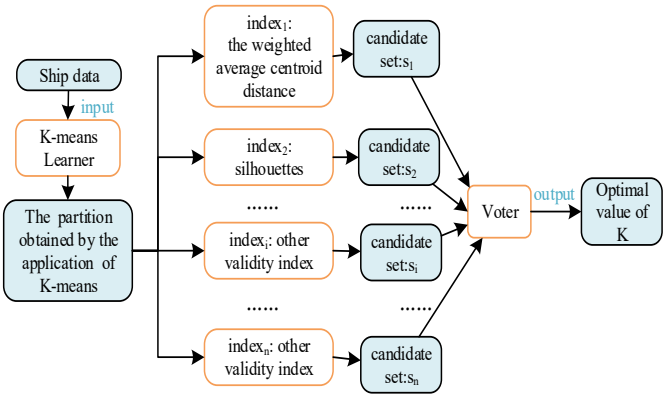

Fig. 7 The structure of the model that determinates the number of clusters automatic based on the voting mechanism

Based on the K value selection model of the voting mechanism, the given data set is used as the input vector of the model, and then a series of clustering results are generated by enumerating different K values;Then, run the ith K-value selection method $\mathrm{Index_i}$ in Figure 7 to obtain $\mathrm{m_i}$ corresponding values ,for different K values, store them from small to large to the set $S_i = \{k_1, k_2, , , k_{m_i}\}$, after run all the $\mathrm{Index_i}(1 \le i \le n)$, we obtain the n candidate sets $S_1, S_2, , , S_n$ of the K value, and then use them as the input of the poller, count each K's frequency P_k, and finally select the highest frequency p_{max} to be the K value of the K-Means algorithm as the number of clustering .

In section II–B, the candidate set S_1={4,5,6,7} is obtained by running the weighted average centroid distance in the cluster and among the clusters as the evaluation $\mathrm{Index_i}$. In section II–C, the candidate set S_2={2,4,8} is obtained by running the silhouettes coefficient as the evaluation $\mathrm{Index_j}$. Taking the two candidate sets as the input of the voting machine, the voting machine counts the frequency of each *K*, get results: $P_{k=4}$=2,$P_{k=2}$=$P_{k=5}$=$P_{k=6}$=$P_{k=8}$=1. Then, select the *K* =4 as the final the clustering number, that is, ships can be divided into four types according to the basic properties of the ship.

IV. An Experimental Analysis on the Rules Learning of Ship Classification Based on Classification Method

For the purpose of verifying the feasibility and validity of the *K* -selection model of K-means based on the voting mechanism, this paper divides the ship tonnage with the optimal value of K (K = 4), and we compares the obtained results with the port's actual business rules. It shows that the K value determined automatically by the model is indeed consistent with the actual business decision.

In order to analyze the results of the classification of the ship tonnage, this paper chooses Naive Bayesian, decision tree, random forest and support vector machine to learn the rules of the results of ship partition. By using the cross-validation method, ship dataset with category labels is divided into training set and test set, including 3637 training data samples and 1213 test data samples. In addition, using the prediction accuracy and calculation time as the classifier evaluation indicators, the performance of each classifier as the table 4.1 shows.

TABLE II THE EVALUATION TABLE FOR DIFFERENT CLASSIFIERS

Classifier	Performance Evaluation Index	
	Prediction Accuracy	*Calculation Time*
Svm_c_rbf	0.624073	1.337787
Decision_tree	1.000000	0.013814
Forest_10	1.000000	0.189129
Forest_100	0.999176	1.771071
Svm_linear	0.755153	1.221822
Bayes	0.957131	0.010961
Svm_c_linear	1.000000	0.014594

As shown in the table II, the classification result of the decision tree is optimal: the highest accuracy (100%), the shortest calculation time (0.013814s). Therefore, the decision tree is used to learn the rules of the ship data. The decision tree chooses the "Gini index" to select the partition attribute in this paper. The final decision tree is as shown in Fig. 8.

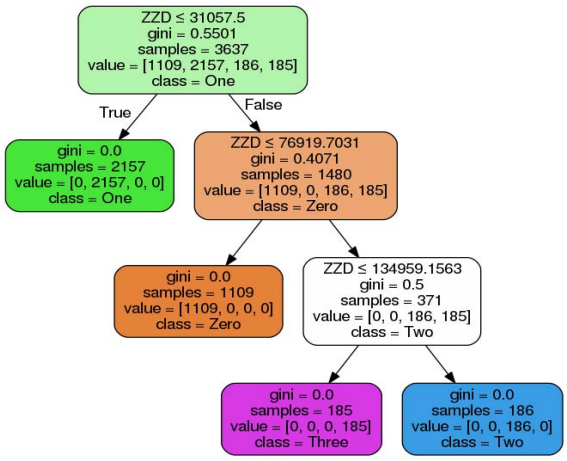

Fig.8 The decision tree based on the Gini index

The experimental decision tree provides some experience and rules for the classification of ship data: according to the load tons, the ship can be divided into four levels, the levels of the load tons from low to high are as shown as follows:

(1) load tons≤310575 ;

(2)310575.5< load tons≤76919.7 ;

(3)76919.7< load tons≤134959.2 ;

(4) load tons>134959.2.

The ship number in the first level is 2157, accounting for 59.31%; that in the second level is 1109, accounting for 30.49%; that in the third level is 185, accounting for 5.09%; and that in the fourth level is 186, accounting for 5.11%.

It can be known that most of the port ships are small tonnage vessels, and large tonnage ships are less. What's more,

the clustering results obtained by dividing the ship according to the basic attributes of the ship are in accordance with the actual business rules. Therefore, the K selection model based on the voting mechanism is feasible and effective.

V. Summary and Prospect

The K-means clustering algorithm is applied to cluster the ship data to find the classification rules of the tonnage grade in this paper. Through the experiment, we found that the K in the application process is involved in the problem of manual intervention. In order to evaluate the merits and demerits of K-means clustering results from several angles, a K selection model based on voting mechanism is proposed to determine the optimal K for obtaining the excellent division. The K selection model based on the voting mechanism automatically obtains the final number of clusters of a given dataset based on the frequency of all K. Through the actual business analysis, it is found that the model can obtain clustering results with high compactness in the cluster, as far as possible between clusters, and even more uniform distribution of samples. By comparing the final clustering results with the actual business rules, it is proved that the model is effective and feasible.

In order to enhance the universality of the model, we can apply the model to other public datasets in the future, and compare the applicability of the model with the experimental results on multiple data sets. In addition, the filter in the model in addition to the usage of clustering validity indexes, can also be other clustering algorithms, such as hierarchical clustering, spectral clustering and so on to obtain K candidate sets.

References

[1] Y.Wang. "Process knowledge mining research and intelligent optimization design in port logistics," Doctoral dissertation, Beijing Jiaotong University, 2014

[2] J.B. Mac Queen. "Some methods for clustering and analysis of multivariate observations," Proc. 5th Berkeley Symp. on Math. Statist. Prob., vol. 1. University of California Press, Berkeley, 1967, pp. 281–297.

[3] B.M. Patil, R.C. Joshi, D. Toshniwal. "Missing Value Imputation Based on K-Mean Clustering with Weighted Distance,". Communications in Computer & Information Science, vol. 94, 2010, pp. 600-609.

[4] T. Zhang, F. Ma. "Improved rough k-means clustering algorithm based on weighted distance measure with Gaussian function," International Journal of Computer Mathematics, 4th ed, vol. 94,2016, pp. 1-17.

[5] K.R. Zalik. "An efficient k-means clustering algorithm," Pattern Recognition Letters, 9th ed, vol. 29, 2008, pp. 1385-1391.

[6] R. H. Thomas, "A majority consensus approach to concurrency control," ACM Trans. Database Syst., vol. 4, 1979, pp. 180-209.

[7] D. K. Gifford, "Weighted voting for replicated data," in Proc.Seventh Symp. Oper. Syst. Principles, 1979, pp. 150-162.

[8] Barbara, Garcia-Molina. "The Reliability of Voting Mechanisms," IEEE Transactions on Computers, 10th ed , vol. C-36 ,2006,pp. 1197-1208.

[9] D. Gondek, T Hofmann. "Non-Redundant Data Clustering ,". Knowledge & Information Systems, 1st ed, vol. 12, 2007,pp. 1-24.

[10] D. Wang, F. Nie, H. Huang. "Feature Selection via Global Redundancy Minimization," IEEE Transactions on Knowledge & Data Engineering, 10th ed, vol. 27,2015,pp. 2743-2755.

[11] O. Arbelaitz, I. Gurrutxaga, J. Muguerza, J.M. Pérez, I. Perona, "An extensive comparative study of cluster validity indices," Pattern Recognit, 1st ed., vol. 46. 2012, pp.243–256.

[12] M.M.-T. Chiang, B. Mirkin, "Intelligent choice of the number of clusters in k-means clustering: an experimental study with different cluster spreads," J. Classif, 1st ed., vol. 27.2010, pp.3–40.

[13] A.K. Jain, "Data clustering: 50 years beyond k-means," Pattern Recognit. Lett. 8th ed., vol. 31,2010, pp.651–666.

[14] D. Steinley, "K-means clustering: a half-century synthesis," Br. J. Math. Stat. Psychol. 1st ed , vol. 59 ,2006, pp.1–34.

[15] P. J. Rousseeuw. "Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis," Computational and Applied Mathematics, vol. 20. 1987, pp.53-65.

[16] R.C.D Amorim, C. Hennig. "Recovering the number of clusters in data sets with noise features using feature rescaling factors," J. Information Sciences An International Journal, 2015, pp. 126-145.

[17] X.J. Fu, L.P. Wang, "Data dimensionality reduction with application to simplifying RBF network structure and improving classification performance," IEEE Trans. System, Man, Cybern, Part B-Cybernetics, vol.33, no.3, pp. 399-409, 2003.

[18] J. Lee; Daewon Lee; "Dynamic Characterization of Cluster Structures for Robust and Inductive Support Vector Clustering," IEEE Trans. Pattern Analysis and Machine Intelligence, vol.28, pp.1869-1874, 2006.

[19] E. Lord, M. Willems, F.J. Lapointe, V. Makarenkov, "Using the stability of objects to determine the number of clusters in datasets," Information Sciences, vol.393, pp.29-46, 2017.

[20] A.V. Kapp, R. Tibshirani, "Are clusters found in one dataset present in another dataset," Biostatistics,vol. 8, 2007,pp. 9–31.

[21] S. Still, W. Bialek, "How many clusters? An information-theoretic perspective," Neural Computation, vol. 16, 2004, pp. 2483–2506.

[22] H. Xiong, J. Wu, J. Chen, "K-means clustering versus validation measures:a data distribution perspective,". IEEE Trans Syst Man Cybern B Cybern, 2006, Vol. 39, pp. 779-784.

[23] M. Bouguessa, S.Wang, H. Sun, "An objective approach to cluster validation," Pattern Recognition Letters, vol. 27, 2006, pp. 1419–1430.

[24] X.L. Xie, G. Beni, "A validitymeasure for fuzzy clustering," IEEE Trans. On Pattern Analysis and Machine Intelligence, vol. 13, 1991,pp. 841–847.

[25] J.C. Dunn, "Well separated clusters and optimal fuzzy partitions," J. Journal of Cybernetics, vol. 4, 1974, pp. 95–104.

[26] P. Lingras, M. Chen, D. Miao, "Rough cluster quality index based on decision theory," IEEE Transactions on Knowledge and Data Engineering, vol. 21, 2009, pp. 1014–1026.

[27] J.Y. Liang, X.W. Zhao, D.Y. Li, F.Y. Cao, C.Y. Dang, "Determining the number of clusters using information entropy for mixed data," Pattern Recognition 45 (2012)