A New Clustering Algorithm Based on Ward_PAM

Hongmei Nie

***Abstract*—Ward algorithm is one of the system clustering methods. The algorithm makes two large classes easily generate a large distance so as to be not easy to be merged, in contrast, it makes two small classes generate a small distance and be easy to be merged. However, the limitation of the method is that it is difficult for the class that has been obtained to be classified again. If we first use Ward method clustering samples, then make each of the obtained classes use PAM algorithm, so that each class can have a chance to be redivided, so get a more detailed clustering effect. In view of this idea, the paper proposes a new clustering method based on Ward and PAM (Ward & PAM algorithm). The proposed method combines the advantages of the two algorithms, which makes the clustering result be more accurate and detailed. Moreover, the paper optimizes the algorithm index formula. Finally, this paper makes a detailed comparison analysis of the experimental results. The experimental result analysis shows that the performance of Ward & PAM algorithm is better than that of Ward algorithm.**

***Index Terms*—Ward algorithm, PAM algorithm, clustering, validity index.**

I. Introduction

Ward algorithm is a hierachical clustering method. The method is designed for sample clustering [1], [2]. Its basic idea is: At first, n samples are respectively considered as n classes, the distance between any two classes is calculated, and the two nearest classes are merged into a new class. Then the distance between the new class and any other class is calculated, the remaining distances are remained unchanged. Again the two nearest classes are merged into a new class. The above steps are repeated, and each time a class is reduced until the desired M classes are obtained [2]–[4].

The method makes two large classes easily generate a large distance so as to be not easy to be merged, in contrast, it makes two small classes generate a small distance and be easy to be merged.

Weighing method is another hierachical clustering method. This method is easy to merge two large classes, in contrast, not easy to merge two small classes [2]. Obviously, Ward method is more reasonable than Weighing method. So, Ward method is a good hierachical clustering method.

However, the limitation of Ward method is that it is difficult for the class that has been obtained to be classified again.

PAM algorithm is a clustering algorithm based on k-medoids [5], [6]. Its main idea is: A cluster is randomly divided into k classes; The representative object of each class is called the center point, and other objects are called non representative objects; The algorithm repeatedly uses non representative objects instead of representative objects to try to find better center points to improve the quality of clustering; In each iteration, all the possible pairs are analyzed, and each pair contains a center point and a non representative object; In each iteration, the optimal object set is the center point set of the next iteration.

The advantage of PAM algorithm is that it is not sensitive to noise and outliers, and can process different types of data.

Imagine: first, we use the Ward algorithm to cluster samples, and then again use the PAM algorithm to redivide each obtained class in order to get more detailed clustering results.

In view of this, the paper proposes a new clustering method based on Ward and PAM (Ward & PAM algorithm). The proposed method combines the advantages of the two algorithms, which makes the clustering result be more accurate and detailed. The experimental result analysis shows that the performance of Ward &PAM is better than that of Ward.

In Section II, the prerequisite knowledge is expounded, in particular, the validity evaluation index formula is optimized. In Section III, the algorithm process of Ward & PAM is described. In Section IV, the simulation experiment is described, and the experimental results are compared and analyzed. In Section V, the conclusion and prospect is proposed.

II. Prerequisite Knowledge

A. Ward Algorithm Basic Idea

A sample C_i (i=1,...,n) represents a class. n_K and n_L are the sample sizes of C_K and C_L, respectively. D_{KL} represents the distance between C_K and C_L. The sum of squared Euclidean distance between each sample of a class and the center of the class is called the sum of squares of deviations. If C_K and C_L are merged into a new class C_M, then the sum of the squares of the deviations of C_K, C_L and C_M are respectively [2], [7]:

$$\begin{aligned} W_K &= \sum_{i\in C_K} (x_i - \bar{x}_K)' (x_i - \bar{x}_K) \\ W_L &= \sum_{i\in C_L} (x_i - \bar{x}_L)' (x_i - \bar{x}_L) \\ W_M &= \sum_{i\in C_M} (x_i - \bar{x}_M)' (x_i - \bar{x}_M) \end{aligned} \tag{1}$$

Let

$$D_{KL}^2 = \frac{n_K n_L}{n_M} (\bar{x}_K - \bar{x}_L)' (\bar{x}_K - \bar{x}_L) \tag{2}$$

Manuscript received October 4, 2015; revised December 10, 2015. This work was supported in part by by the Foundation Name under Grant No. LY13F020016.

Hongmei Nie is with Zhejiang Normal University, Jinhua, China (e-mail: nhm@zjnu.cn).

Formula (2) is called the square distance between C_K and C_L. Then the distance recursion formula is [2]:

$$D_{MJ}^2 = \frac{n_J + n_K}{n_J + n_M} D_{KJ}^2 + \frac{n_J + n_L}{n_J + n_M} D_{LJ}^2 - \frac{n_J}{n_J + n_M} D_{KL}^2 \quad (3)$$

The basic idea of Ward algorithm is described as follows:

At first, *n* samples are respectively considered as n classes; then by Formula (2), the distance between any two classes is calculated, and the two nearest classes are merged into a new class; By Formula (3), the distance between the new class and any other class is calculated, the remaining distances are remained unchanged, and again the two nearest classes are merged into a new class. The above steps are repeated, and each time a class is reduced until the desired *M* classes are obtained.

B. PAM Algorithm Basic Idea

The basic idea of PAM algorithm is described as follows:

First a representative object is selected for each class. That is the center point O_i (i=1,…,k). And then, at every step, the center point O_i is replaced by the non central point O_h, which leads to the improvement of clustering quality. O_j represents any non central object, d (O_j, O_m) represents the Euclidean distance between the object O_j and O_m , and C_{jih} represents the cost of O_i being replaced by O_h. For each situation, C_{jih} is respectively discussed as follows:

Let O_i be the center point of the class CO_i, and O_g ($i \neq g$) is another center point that does not belong to CO_i. If $O_j \in CO_i$, and d (O_j, O_h) > d (O_j, O_g), then $C_{jih} = d$ (O_j, O_g) - d (O_j, O_i);

Let O_i be the center point of the class CO_i, and O_g ($i \neq g$) is another center point that does not belong to CO_i. If $O_j \in CO_i$, and d (O_j, O_h) < d (O_j, O_g), then $C_{jih} = d$ (O_j, O_h) - d (O_j, O_i);

Let O_i be the center point of the class CO_i, O_g be the center point of the class CO_g, and $CO_i \neq CO_g$. If $O_j \in CO_g$, and d (O_j, O_h) > d (O_j, O_g), then $C_{jih} = d$ (O_j, O_g) - d (O_j, O_g) = 0;

Let O_i be the center point of the class CO_i, O_g be the center point of the class CO_g, and $CO_i \neq CO_g$. If $O_j \in CO_g$, and d (O_j, O_h) < d (O_j, O_g), then $C_{jih} = d$ (O_j, O_h) - d (O_j, O_g).

To sum up, the total cost of O_i being replaced by O_h is: $TC_{ih} = \sum C_{jih}$. If TC_{ih} is less than zero, then O_i is replaced by O_h. Repeat the above steps until we find the most similar representative object for each non central point.

C. Algorithm Validity Index

The algorithm validity index DB(Davies-Bouldin) is given as follows [8]:

$$DB = \frac{1}{k} \sum_{i=1}^{k} R_i \quad (4)$$

where

$$R_i = \max_{\substack{j=1,\ldots,m \\ i \neq j}} R_{ij} \quad (5)$$

$$R_{ij} = \frac{S(C_i) + S(C_j)}{D(C_i, C_j)} \quad (6)$$

$$S(C_i) = \frac{1}{|C_i|} \sum_{x_i \in C_i} \| x_i - \bar{x}_i \| \quad (7)$$

$$D(C_i, C_j) = \| \bar{x}_i - \bar{x}_j \| \quad (8)$$

In the above formulas, k represents the total number of all classes, m is the label of a class, x_i is the sample of a class C_i , and $\bar{x}_i$ is the sample mean of a class C_i. In a class C_i, $S(C_i)$ represents the mean distance between all the samples and the center $\bar{x}_i$, $| C_i |$ represents the sample number of a class C_i, and $D(C_i, C_j)$ represents the distance between the class C_i and C_j. The above formula(6) shows that the small numerator and the bigger denominator make the clustering effect be better.

But the author believes that the formula (5) only consider the special factor, that is the maximum value of R_{ij}, which will lead to the validity index with error. For this reason, the author modifies the above formulas, and uses the correlation coefficients to replace all the distances. The modified formulas are described as follows:

$$DB = \frac{\frac{1}{k} \sum_{i=1}^{k} S(C_i)}{\frac{1}{k^2} \left(\sum_{i=1}^{k} \sum_{j=1}^{k} r(C_i, C_j) + k \right)} \quad (9)$$

where

$$S(C_i) = \frac{1}{|C_i|} \sum r(x_i, \bar{x}_i) \quad (10)$$

$$r(x_i, \bar{x}_i) = \frac{x_i \cdot \bar{x}_i}{\| x_i \| \cdot \| \bar{x}_i \|} \quad (11)$$

$$r(C_i, C_j) = \frac{\bar{x}_i \cdot \bar{x}_j}{\| \bar{x}_i \| \cdot \| \bar{x}_j \|} \quad (12)$$

In (11), $r(x_i, \bar{x}_i)$ represents the correlation coefficient between x_i and $\bar{x}_i$. In (12), $r(C_i, C_j)$ represents the correlation coefficient between C_i and C_j. Obviously, the modified formulas take into account the overall situation, so they are more reasonable. In (9), DB can be used as the effective evaluation index for a clustering algorithm. From the formulas, we can see that the bigger the index DB is, the better the clustering effect is.

III. IMPROVED WARD AND PAM ALGORITHM

The improved Ward & PAM algorithm is described as follows:

Step 1: Let each sample be a class C_i (i=1,…,n). According to the formula (2), calculate the distance between any two classes, and get the distance square matrix $D(0)=(D_{KL}^2)$, K, L=1,…,n. $D(0)$ is a symmetric matrix;

Step 2: Select the smallest element in $D(0)$, and let it be D_{KL}, then merge C_K and C_L into a new class C_M;

Step 3: According to the formula (3), calculate the distance

D^2_{MJ} between the new class C_M and any other class C_J, and remain the remaining distances in $D(0)$ unchanged. So get the new distance matrix $D(1)$;

Step 4: On the $D(1)$, repeat step 2 and step 3, and get the matrix $D(2)$. So repeat until we get M classes.

Step 5: According to Ward method, we obtain M classes. For each of the obtained classes, use PAM algorithm. That is, for each class , choose a representative object (a center point O_i, i=1,…,k);

Step 6: Select a non central point O_h;

Step 7: Calculate the total cost TC_{ih} of each O_i (i=1,…,k) being replaced by O_h;

Step 8: Select the pair (O_i, O_h) corresponding min(TC_{ih}) (i=1,…,k). If min(TC_{ih}) is less than zero, then O_i is replaced by O_h, and return to step 5;

Step 9: Otherwise, for each non central point, find the most similar representative object.

IV. Simulation Experiment

This paper selects the two data sets in "2014 China Statistical Yearbook" as the test data sets. The first data set is the per capita cash consumption data of urban residents in the different regions. The second data set is the data of the general higher schools in the different regions. On the Matlab7.0.1 platform, we implement the testing of Ward and Ward & PAM.

A. Experiment

The sample of *Experiment* 1 is the data of the per capita cash consumption expenditure of urban residents in 31 regions of China. The attributes of the data set are: consumption expenditure in cash, food, clothing, housing, household equipment and supplies, transportation and communication, education and entertainment, medical care, other.

As shown in Table I, the 31 regions of China are divided into five classes by using the Ward algorithm.

TABLE I: Ward Test Results

Classes	regions
I	Beijing, Shanghai
II	Zhejiang, Guangdong, Jiangsu, Tianjin, Fujian
III	Liaoning, Shandong, Chongqing, Inner Mongolia
IV	Anhui, Sichuan, Shaanxi, Hubei, Hainan, Guangxi, Yunnan, Hunan, Ningxia, Xinjiang, Henan, Jilin
V	Gansu, Guizhou, Jiangxi, Hebei, Qinghai, Shanxi, Tibet, Heilongjiang

As shown in Table II, the fourth class and fifth class (Set class parameter k=2) are redivided, and the 31 regions of China are divided into seven classes by using the Ward & PAM algorithm.

The sample of *Experiment* 2 is the data of the general higher school situation in the 31 regions of China. The data is the statistic on the number of the attributes. The attributes of the data set are: school, professor, associate professor, lecturer, assistant lecturer, no title teacher, administration staff, counselor, and handyman.

As shown in Table III, the 31 regions of China are divided into six classes by using the Ward algorithm

TABLE II: Ward and PAM Test Results

Classes	regions
I	Beijing, Shanghai
II	Zhejiang, Guangdong, Jiangsu, Tianjin, Fujian
III	Liaoning, Shandong, Chongqing, Inner Mongolia
IV	Shaanxi, Sichuan, Hunan, Anhui, Hubei, Hainan, Guangxi
V	Jilin, Yunnan, Ningxia, Xinjiang, Henan
VI	Gansu, Guizhou, Jiangxi, Hebei, Qinghai, Shanxi, Heilongjiang
VII	Tibet

TABLE III: Ward Test Results

Classes	regions
I	Tibet, Qinghai, Hainan, Ningxia, Inner Mongolia, Guizhou, Gansu, Xinjiang
II	Guangxi, Chongqing, Yunnan, Tianjin, Shanxi, Fujian, Jilin, Anhui, Jiangxi, Heilongjiang, Shanghai
III	Jiangsu, Shandong
IV	Henan, Sichuan, Hubei, Guangdong
V	Hebei, Liaoning, Hunan, Shaanxi, Zhejiang
VI	Beijing

As shown in Table IV, the first class and second class (Set class parameter k=2) are redivided, and the 31 regions of China are divided into eight classes by using the Ward & PAM algorithm.

TABLE IV: Ward and PAM test results

Classes	regions
I	Tibet, Qinghai, Hainan, Ningxia, Xinjiang
II	Inner Mongolia, Guizhou, Gansu
III	Guangxi, Chongqing, Yunnan, Tianjin, Shanxi, Fujian, Jilin
IV	Anhui, Jiangxi, Heilongjiang, Shanghai
V	Jiangsu, Shandong
VI	Henan, Sichuan, Hubei, Guangdong
VII	Hebei, Liaoning, Hunan, Shaanxi, Zhejiang
VIII	Beijing

B. Comparative Analysis Combined with the Actual Consumption of Urban Residents in Various Regions

By combining the actual consumption of urban residents in different regions of China [9], [10], the author makes a comparative analysis of the two algorithms.

The first consumption group includes Beijing and Shanghai. They are the two centers of China's economy, transportation, science, technology, industry, finance, exhibition and shipping. The regional advantage of these two areas is obvious, so the residents must spend more money.

The second group includes Zhejiang, Guangdong, Jiangsu, Tianjin and Fujian. These regions are the most developed regions of China's economy, so the intensity of the consumer spending of the residents is bound to be the biggest.

The third consumer group includes Liaoning, Shandong, Chongqing and Inner Mongolia. In recent years, these regions have maintained rapid economic growth, thus effectively stimulating the consumption level of the residents in these regions.

In terms of the above three consumption groups, Table I and Table II draw the same clustering result, so the result is reasonable.

In recent years, Anhui, Sichuan, Shaanxi, Hubei, Hainan, Guangxi and Hunan have maintained relatively rapid economic growth rate, the income of the residents has increased year by year, and the consumer awareness of the residents has continued to increase. So, these regions belonging to the fourth consumer group is more reasonable. In Table I, Jilin, Yunnan, Ningxia, Xinjiang and Henan are divided into the fourth consumer group. This clustering result seems reasonable. But considering the actual situation, we can find some errors in this clustering. Although Jilin, Yunnan, Ningxia and Xinjiang have maintained the economic growth, but the growth rates haven't been high, the industrial infrastructure and industrial system haven't been very perfect. So these regions are still under developed areas, and the overall consumption level is not high. In recent years, Henan has maintained a rapid economic development, but its regional developments have been not balanced. So it has affected the average consumption level of the residents. So, Jilin, Yunnan, Ningxia, Xinjiang and Henan should belong to the fifth consumer group. Therefore, it can be seen that the result of Table II is more reasonable than that of Table I.

Gansu, Guizhou, Jiangxi, Qinghai, Shanxi and Heilongjiang in recent years have maintained steady economic growth. But these regions have been located in remote areas, the transportations haven't been convenient, economic infrastructures haven't been perfect, the people's developement awareness hasn't been strong, and the social security systems haven't been good enough. So many reasons lead to the relative backwardness of the income and consumption of the residents. Hebei has developed rapidly economy. But its regional developments haven't been balanced, so the average consumption growth of the residents has been restricted. Thus, Gansu, Guizhou, Jiangxi, Qinghai, Shanxi, Heilongjiang and Hebei should belong to the sixth consumer group. Tibet has been located in the remote areas, its culture, economic infrastructure and transportation have been relatively backward, the incomes of the residents haven't been balanced, and people's development ideas and consumption consciousness have been relatively indifferent. Many reasons lead to Tibet is a area which consumption level is the most backward. Therefore, it can be seen that the result of Table II is more reasonable than that of Table I.

Through the above analysis, we can know that the result of the Ward & PAM algorithm is more reasonable than that of the Ward algorithm.

C. Comparative Analysis Combined with the Actual General Higher School Situation in Various Regions

The second experiment is about the analysis of the general higher school situation in the 31 regions of China. The data mainly shows the distributions of general higher school situation in various regions. Table III and Table IV respectively show the results of the two algorithms. Table III shows that the clustering of the third group, fourth group, fifth group and sixth group is as same as that of the corresponding groups of Table IV. This clustering reflects the educational form of each region is in line with local education history and geographical environment. With the economic development in recent years, the education resources have been integrated, and the new discipline categories have been developed, therefore, the two tables also reflect the educational form of each region is consistent with the local economic development.

The result of the first group of Table III is not reasonable enough. Although the general higher education foundation of Inner Mongolia, Guizhou and Gansu in the history was weak, still had certain education foundation. And with the economic developments, the general higher education situation of these regions is also improving year by year [11]. Thus, these regions should belong to the same group. And for Tibet, Qinghai, Hainan, Ningxia and Xinjiang, in the history, the general higher education foundation of these regions was almost zero. With the economic development, the general higher education of these regions has been developed. But because of historical reasons, the general higher education in these regions is still relatively backward [11]. So, these regions should belong to another group. Therefore, the result of Table IV is more detailed than that of the Table III.

The result of the second group of Table III is basically reasonable. But by the analysis of the actual situation, we can find some deviations in the result. On the whole, the average education development levels of the second groups are same. But Anhui, Jiangxi, Heilongjiang and Shanghai have always had prominent educational institutions [11], so the four regions should be divided into one category. Therefore, the result of Table IV is more detailed than that of Table III.

The above analysis shows that the result of the Ward & PAM algorithm is more reasonable than that of the Ward algorithm.

D. Comparative Analysis of Validity Index

In the first experiment, DB1 and DB2 respectively represent the validity indexes of the two methods. By the optimized validity index formula, we calculate the value of DB1 and DB2. That is: DB1 ≈ 0.8425, DB2 ≈ 0.8785, and DB1< DB2.

In the second experiment, DB3 and DB4 respectively represent the validity indexes of the two methods. By the optimized validity index formula, DB1 and DB2 are calculated as follows: DB3 ≈ 0.8513, DB4 ≈ 0.8781,and DB3< DB4.

The two results show that the index of Ward & PAM algorithm is bigger than that of Ward algorithm. Therefore, the performance of Ward & PAM algorithm is better than that of Ward algorithm.

V. Conclusion

This paper proposes a new clustering method based on Ward and PAM. The algorithm optimizes Ward algorithm by using PAM algorithm. Because the algorithm combines the advantages of PAM and Ward, the clustering effect is more

accurate and detailed.

The proposed new algorithm can not only deal with the general data sets, but also provides a feature selection method for text classification.

It is a common phenomenon for us to deal with general data sets. However, when the sample number is too large, the cost of Ward & PAM is very large. So for dealing with large data sets, we further need to improve the algorithm. In view of CLARA algorithm being suitable for processing large data sets [12], [13], the next step of our research is to develop a clustering method based on Ward & PAM and CLARA.

REFERENCES

[1] H. Zhao, L.-H. Zhu, and D. Liu, "Application of ward system clustering method in multi variable sampling technique," *Statistics and Decision*, pp. 68-69, 2006.

[2] X.-M. Wang, *Applied Multivariate Analysis*, Shanghai Finance University Press, pp. 163-166, 2014.

[3] B. Wu, H.-G. Xu, and W.-H. Zhang, "Multivariate statistical analysis of road traffic accidents based on Ward clustering method," *Journal of Heilongjiang Institute of Technology*, vol. 23, pp. 4-6, 2009.

[4] X.-T. Yang, H.-T. Yang, and Z.-B. Xu, "Research on reservoir classification evaluation based on principal factor analysis and WARD clustering method," *Acta Geologica Sichuan*, vol. 33, pp. 113-115, 2013.

[5] D. Klein, S. Kamvar, and C. Manning, "From instance-level con-straints to space-level constraints: making the most of priorknowledge in data clustering," in *Proc. the 19th International Conference on Machine Leaming*, Stan-ford, CA, USA, pp. 307-314, 2002.

[6] M.-X. Duan and K. Sun, "RBF neural network design based on PAM clustering method," *Journal of Shenyang Normal University*, vol. 27, pp. 440-443, 2009.

[7] W.-D. Li, *Applied Multivariate Analysis*, Peking University Press, 2008, pp. 127-128.

[8] S.-H. Li and H.-M. Zhao, "Clustering validity analysis," *Metallurgical Automation*, pp. 336-338, 2004.

[9] X.-Y. Zhang and W.-Q. Li, "The influence factors and the effect of the regional differences in the consumption of urban residents," *Consumption Economy*, vo1. 27, no. 6, pp. 37-40, December 2011.

[10] A.-J. Sun, "Study on regional characteristics of urban residents' consumption," *Consumption Economy*, vo1. 26, no. 5, pp. 7-11, October 2010.

[11] L. Zhao, J.-H. Shi, P. Wang, W. Wang, and T. Xu, "Analysis on the types and regional differences of higher education quality," Education Research of Tsinghua University, vo1.33, no. 5, pp. 1-12, October 2012.

[12] G.-F. Zhao and G.-Q. Qu, "Analysis and implementation of CLARA algorithm in clustering analysis," *Journal of Shandong University of Technology*, vol. 20, pp. 45-48, 2006.

[13] L.-G. Tong, J.-X. Xie, and M.-L. Gao, "The implementation of CLARA algorithm on Protein sequences clustering," *Microcomputer Information*, vol. 26, pp. 231-233, 2010.

Hongmei Nie was born in Sichuan province, China, in 1968. She is an associate professor at the Zhejiang Normal University. Her research interests include data mining, computer application, and artificial intelligence. She published many papers in domestic and international academic journals and conferences, and participated in several national or provincial scientific research projects.