International Journal of Medical Science and Clinical Inventions 4(3): 2768-2773, 2017
DOI:10.18535/ijmsci/ v4i3.8 **ICV 2015: 52.82**
e-ISSN: 2348-991X , p-ISSN: 2454-9576

Research Article

Comparison of EM and Two-Step Cluster Method for Mixed Data: An Application

Özge Pasin[1], Handan Ankarali[2]

[1]Istanbul University Biostatistics Department, Turkey
[2]Duzce University Biostatistics Department, Turkey

ABSTRACT:
There have been more than 50 type clustering algorithms developed for getting meaningful information from big datasets and grouping individuals according to their characteristics.
In actual researches, it is often seen that data involves all types of variables. In this case, it is very important to select appropriate clustering algorithm according to different data types. In this study, we will provide information about EM(Expectation Maximization),Two–Step Clustering methods which are developed in recent years and one of the best methods for data sets containing mixed types of variables. And the second aim is to compare the methods by producing a data set from health field information.These algorithms are generally recommended for large data sets but there are also used n medium-sized data sets. Medium- sized data sets are more often in actual researches.Therefore, fifty people for control group and fifty people for patients that have polycystic over syndrome were taken to the study. Totally nineteen variables were measured from these subjects and thirteen of them were quantitative, six of them were qualitative.Clusters were obtained by EM and Two-Step cluster methods.To evaluate the relationships between the clusters obtained from algorithms and actually known patient, control groups were analyzed by Kappa coefficient. It was found that EM clustering algorithm has highest compliance coefficient comparing with Two-Step cluster(Kappa=0,740;p<0,001) and it was seen EM method was a better algorithm for finding both patients and controls.
As a result, we can say that researchers may have successful results for classifying diseases by appropriate clustering methods.

Key Words: **Clustering, Data Mining, EM, Polycystic over syndrome, Two-Step Clustering**

1.Introduction

Clustering is a process for multivariate data analysis. This analysis is an important human activity for distinguishing. It partitions a set of data objects into subsets and each subset is a cluster. The objects that included in the same cluster have similar features and similar distances from cluster centers. Cluster analysis is the main technique for data mining science. It can use in all science field such as web search, biology, education, engineering, health, medicine etc. Also in health researches you can use clustering for analysis of regional disease, personnel management, timing of ambulance transport services, classification of physiological states, detection of tumors by the help of MR and ultrasound, determining the density of traffic accidents, diagnosis of disease, determining of the different morphology of the heart sound, distribution of health units and these examples can also be increased. Cluster analysis can be also used for obtaining homogeny groups as preliminary statistical analysis (Ferligoj,1983;Fraley 2005).

There are lots of clustering algorithms such as Hierarchical Clustering Methods. Density Based Clustering Methods, Partitioning Clustering Methods, Grid-Based Clustering Methods, Categorical Clustering Methods, Model-Based Clustering Methods, Hybrid Clustering Methods, Fuzzy Clustering Methods. These are eight main cluster groups. The choice of a suitable clustering algorithm depends on the clustering objects and clustering task.

A good clustering algorithm should have some features. It should cluster both big data and small data sets. Also, it should have to deal with mixed data such as binary, ordinal, nominal or numerical attributes. The other feature of a good clustering is discovering clusters with arbitrary shape. A cluster could be any of shape and the other issue is, in health studies there are lots of missing observations or unknown data. The algorithm should be deal with these observations and noisy data (Han,2006).

When clustering objects, some algorithms need a knowledge for determining input parameters like a number of clusters and analysis is very sensitive to this parameter. So a good method should minimize these input parameters that specified by the user. The results of this algorithm should be usable, interpretable. And the last feature of a good algorithm is a capability of high dimensionality data (Han,2006).

In our study, we will give information about two clustering methods that used in this study named as Expectation Maximation algorithm and Two Step cluster analysis that located in the above methods. And for the second aim of this study, we will show and discuss results about comparing these methods. for the application. So in the next section, we are going to focus on these methods.

2. Material and Methods

2.1. Expectation Maximization (EM) Clustering Algorithm

EM clustering algorithm is an unsupervised method. It is used to estimate the density of data points. It is a model based algorithm. In this method, each cluster represents mathematically by a probability distribution. EM clustering algorithms first start to make predictions about the parameters including covariance. Then there are two steps including expected step (expectations) and maximization step. The name

$$p(z_{nk}=1|x_n,\mu_k,\Sigma_k)=\frac{\pi_k N(x_n|\mu_k,\Sigma_k)}{\sum_{j=1}^{K}\pi_j N(x_n|\mu_j,\Sigma_j)}=\gamma(z_{nk}) \qquad (1)$$

In M step, $Q(\Theta,\Theta^{(t)})$ should be maximized. The expected loglikelihood of complete data can be calculated by the following equation under the independence assumption.

$$Q(\Theta,\Theta^{(t)})=\sum_{n=1}^{N}\sum_{k=1}^{K}\gamma(z_{nk})^{(t)}log\pi_k N(x_n|\mu_k,\Sigma_k) \qquad (2)$$

Initial values are selected for $\{\mu_k\}$ mean vector. Then two stages are repeated until obtaining a stable result. This algorithm is based on some intensive basic statistics techniques and it is robust to noisy data. It can be used for high dimensional data. The steps of EM clustering is simple and easy to understand.It has the ability to estimate missing observations in the data.It has less cost than other clustering algorithms (Aggarwal 2014;Han 2006).

2.2. Two -Step Clustering Algorithm

Two-Step clustering algorithm combines both hierarchical and partitioning methods. Two- Step clustering method utilizes a two step approach similar to BIRCH (Zhang, 1996). Two-Step method involves two steps including Pre-clustering and Clustering steps.

Pre-clustering step scans the data record one by one and decides whether the current record can be added to one of the previously formed clusters or it starts a new cluster based on distance criterion. The method uses two types of distance measuring Euclidian and loglikelihood distance. Euclidian distance can be used for categorical variables but loglikelihood measure can be used for both categorical and numerical variables (Banfield, 1993; SPSS,2001).

Pre-clustering step is similar progress like BIRCH algorithm. It uses Clustering Feature (CF) for clustering. In CF there are nodes and these nodes have a number of entries. In this step, it is investigated that what is the nearest leaf entry in leaf nodes. If this leaf entry is within thethreshold distance that determined initially, it is included into the nearest leaf entry. Otherwise, a new value is generated for the leaf node (SPSS 2001; Zhang 1996).

In clustering step subclusters are used obtained from pre-clustering step as an input and then they are grouped in the desired number of clusters. Also in this method, there is no need to specify an input parameter like a number of clusters. Because method did this automatically by the help of BIC and AIC information criterions. The initial estimation of a number of clusters is calculated easily with this indicator. An E step comes from the fact that there is only need to compute expected sufficient statistics. The name M step comes from, model reestimation. It maximizes the expected log likelihood of the data (Aggarwal, 2014; Han 2006).

EM algorithm is a popular iterative method to find the hidden variables probability of the ML and MAP estimates. In E step, the hidden parameters (z_{nk}) posterior probabilities are calculated. The following equation is obtained using Bayesian theorem (Aggarwal 2014).

important advantage of this method is, it can be used for mixed data types like ordinal, nominal or numeric. And it can work well with big datasets that may contain million or billion of objects with a short time. Even if data contain outliers or normality assumption is not met, Two-Step clustering method gives appropriate results. But is not usable for data sets that contain a missing value. So before making analysis with this method, data should be examined and missing values must be evaluated (Schiopu 2010; SPSS 2001).

3. Application and Statistical Analysis

In our country and in all word polycystic over syndrome disease is the most common endocrine disorder disease in recent years for women. It has lots of risk factors such as obesity, diabetes, menstrual disorders, skin problems, age, body mass index etc. Also some genetic factors. Polycystic over syndrome disease's etiopathogenesis is not clearly known for this available treatment options is usually symptomatic currently (Stein 1935).So we want to ensure a little contribution to this lack by cluster analysis that used new, usable, good methods. The data used in our study was about patients that have polycystic over syndrome and we generated values by using descriptive statistics obtained from literature with a simulation study. 100 individual measures were obtained. We wanted to investigate that what is the risk factors of polycystic over syndrome and it is the answer of how to discriminate the groups. Our main question in this study is which method (EM or Two-step clustering) best split the groups by looking actual groups. Also, we know where each person is included to control or polycystic over syndrome patients. So we have two groups included control and patients. Then there are some variables in the below that used in this study for analysis.

- Age, body mass index, waist-hip ratio
- Duration of menses, Triglycerides, HDL, LDL, FSH, LH
- Prolactin, Estradiol, Testosterone, TSH.
- Disorder of ovulation (Yes, No)
- Insulin Resistance (Yes, No)
- Disorder of menstrual (Yes, No)
- Increase of pubescence (Yes, No)
- Acne Problem (Yes, No)

- Lubrication of skin (Yes, No)

Data have both numerical and categorical variables and we used these variables to look how successful grouping because we know actual two groups.

For statistic analysis, numerical variables descriptive statistics were given as mean, standard deviation, minimum and maximum. For categorical variables statistics were given as frequence and percentage. Clustering process is made by EM and Two-Step clustering methods. Concordance of clustering algorithms were evaluated with Kappa statistics. The statistical significance level was 0,05 and WEKA and SPSS (ver.21) was utilized for the analysis.

4. Results

All numerical variables descriptive statistics were given as mean, standard deviation, minimum and maximum in Table 1.

Table 1. Descriptive statistics for numerical variables

Variables	Mean	Std. Deviation	Minimum	Maximum
Age	23,94	3,876	17,00	32,00
Body Mass Index	25,88	4,033	18,51	39,00
Waist-hip ratio	0,84	0,071	0,60	0,99
Duration of menses	42,51	30,739	18,00	180,00
Triglycerides	100.64	56.388	35.00	323.00
HDL	51.09	12.223	30.00	88.00
LDL	93.61	21.371	54.00	150.00
FSH	5,66	1,789	2,00	9,20
LH	5,87	2,571	1,00	13,00
Prolactin	11,98	6,685	1,00	45,00
Estradiol	70,15	41,467	10,00	217,00
Testosterone	50,72	16,481	16,00	92,00
TSH	2,34	0,850	1,00	4,50

Considering Table 2 results, you can see frequences for categorical variables. 44% of people who participated in the study had ovulation disorder, 39% had insulin resistance, 47% had menstrual problems, 39% pubescence increase, 49% had acne problem and 47% had skin lubrication.

Table 2. The distribution of categorical variables

Variables	Percentage of Yes Answers
Disorder of ovulation	%44 (44 person)
Insulin Resistance	%39 (39 person)
Disorder of menstrual	%47 (47 person)
Increase of pubescence	%39 (39 person)
Acne Problem	%49 (49 person)
Lubrication of skin	%47 (47 person)

Considering age, body mass index, waist-hip ratio, duration of menses, Triglycerides, HDL, LDL, FSH, LH, Prolactin, Estradiol, Testosterone, TSH, Disorder of ovulation, insulin resistance, disorder of menstrual, ,increase of pubescence, acne problem and lubrication of skin variables in the data, EM and Two-Step Clustering methods were applied.

According to Two-Step clustering, we obtained Table 3, 4 and 5. We used for determining the number of clusters by examining BIC criteria and the results were obtained in Table 3. This table shows various cluster members obtained for determining suitable cluster number in grouping data by looking the similarities. We found that data should be separated into two clusters since its ratio distances are the largest.

Table 3. Determining number of clusters

Number of Clusters	Schwarz's Bayesian Criterion (BIC)	Ratio of Distance Measures
1	1861,754	
2	**1645,077**	**3,770**
3	1695,886	1,011
4	1747,792	1,568
5	1834,278	1,230
6	1932,158	1,171
7	2037,279	1,004
8	2142,565	1,053
9	2249,986	1,180
10	2363,512	1,064
11	2479,074	1,031
12	2595,599	1,092
13	2714,733	1,067
14	2835,630	1,050
15	2957,781	1,193

In Table 4, the relationship between Two-Step clustering and actual groups was evaluated by a crosstab. In Two-Step cluster analysis results, we found 3 people were patient while their actual group was control and 20 people were control while their actual group was patient. So 23 people clusters were obtained wrongly. But 77 people were included correctly to their groups. The proportion of clustering controls correctly was 94%, and for the patient the proportion was 60%. So Two-Step methods found the controls more rightly comparing with patients.

Table 4. Relationship between Two-Step Cluster method and actual groups

Two Step Clustering Results			Actual Groups		Total
			Control	Patient	
TwoStep Cluster Method	Control	Count	47	20	67
		% within Two-Step Cluster Method	70,1	29,9	100
		% within Actual Groups	**94**	40	67
	Patient	Count	3	30	33
		% within Two-Step Cluster Method	9,1	90,9	100
		% within Actual Groups	6	**60**	33
Total		Count	50	50	100

Concordance of the clustering results for Two Step clustering was investigated with Kappa statistics and the results were shown in Table 5. According to Table 5, there was significant harmony among Two-Step clustering results and actual groups. But kappa coefficient was quite small as you can see in this table (Kappa=0,540).

Table 5. Kappa coefficient between groups obtained from Two-Step Clustering and Actual Groups

		Value	Asymptotic Standardized Error	Approximate T	Approximate Significance
Measure of Agreement	Kappa	**0,540**	0,079	5,742	<0,001

In Table 6 the relationship between EM method and actual groups was evaluated by a cross table. We found that out of the 41,

who were patient in terms of EM clustering result, 39 were really patient. So in this method, the success of finding real patients were 78% , the success of finding real control were 96%. The proportion of correctly clustering in terms of both patients and controls increased when comparing with Two-Step clustering method results.

Table 6. Relationship between Expectation Maximization algorithm and actual groups

EM Clustering Results			Actual Groups		Total
			Control	Patient	
Expectation Maximization	Control	Count	48	11	59
		% within Em	81,4	18,6	100
		% within Actual Groups	**96**	22	59
	Patient	Count	2	39	41
		% within Em	4,9	95,1	100
		% within Actual Groups	4,0	**78**	41
Total		Count	50	50	100

Table 7 was obtained by evaluating the relationship between EM and actual groups. There was a significant harmony between these results. Also kappa coefficient was higher than Two-Step analysis results.

Table 7. Kappa coefficient between groups obtained from Expectation Maximization algorithm and Actual Groups

		Value	Asymptotic Standardized Error	Approximate T	Approximate Significance
Measure of Agreement	Kappa	**0,740**	0,066	7,523	<0,001

5. Discussion

Data mining results have been developed for a large number of variables and data sets that contain a large number of individuals. Usually, it is used for classifying individuals or variables based on the similarity between individuals and variables and there are lots of algorithms for this (Kob, 2005). It is important to select the correct clustering method for applications and these selection steps are depends on the properties of variables and sample size. Many studies that use clustering algorithms in health studies. But we think that these studies should be increased by researches. There are lots of reasons that we should increase the usage of clustering in health researches. For example for diagnosis of disease, distribution of health units, personnel management in hospitals, detection of tumors, eliminate the subjective opinion of doctors about patients that have unclear symptoms or determining the risk factors for a disease etc.

In our study, we investigated Polycystic over syndrome risk factors. We clustered Polycystic over syndrome patients and controls by looking some variables including both numerical and categorical type. We used EM and Two-Step Cluster Methods and we compared these two methods results with each other. It was found that EM clustering algorithm has highest compliance coefficient comparing with Two-Step cluster (Kappa=0,740; p<0,001). It was seen that compared with Two-Step cluster algorithm, EM method was a better algorithm for finding both patients and controls. So EM algorithm is better than Two-Step analysis for our application data. But this result is not enough. These results should be considered as clinically. Also in some studies, finding patients is less important than controls but in some studies, it is the reverse. Results should be investigated depends on this assessment.

We could not get available results when we compare parallel studies in the literature that compared EM and Two-Step clustering algorithms. But we observed that EM clustering algorithm was compared with other clustering methods in most research. For example, Zheng et al compared EM, farthest first and K-means clustering algorithms in a data set. They found that EM algorithm was superior to other methods for all criteria. Also, they have determined that EM algorithm had a smaller standard deviation from K-means and farthest first clustering methods for all data sets (Zheng 2005).

In 2008, Osama Abbas compared different clustering algorithms and he has concluded that EM algorithms had better performance from hierarchical clustering methods. In addition, he emphasized that EM and K-means methods produced very good results for large databases. (Abbas, 2008). In 2012 Sharma and colleagues compared algorithms that used in WEKA program and they found EM clustering algorithm is very useful for real data sets (Sharma, 2012).

Kakkar and Parashar compared K-means, hierarchical methods, EM and density based algorithms that used in WEKA in 2014. As a result of their study, they observed that K-means clustering algorithm gave faster results than hierarchical and EM algorithm (Kakkar 2014).

Goyal concluded that the best methods were EM and K-means algorithm from COBWEB, DBSCAN and farthest first algorithms that used in WEKA by applying the datasets in 2014 (Goyal, 2014).

Jung et al., compared K-means and EM clustering methods in 2015. The results of their study shows that, K-means algorithms accuracy was higher than EM clustering. But they determined that K-means algorithm took more time than EM (Jung, 2014).

As a result, we can say that researches can have errors, if they reach a definitive conclusion that this gives better results in the dataset. Clustering algorithms should be reviewed by taking account clinical information, evaluating methods criteria, assumptions, conditions of use, advantages and disadvantages as a whole. Statistical methods must be in support of the clinical findings for using easily and getting correct results in the application.

We should not forget that researchers can obtain successful results for classifying diseases by appropriate clustering methods. If correct method is used, health policy will be developed and individuals who have high risks will be determined. When high-risk individuals identified, necessary precautions will be taken in the future. So a basic clustering algorithm application can improve and make differences in the health area. A basic clustering algorithm can improve public's quality of life and can increase life expectancy of public.

The limitation of this study is to compare two cluster methods by using a single set of data. A simulation study will be planned for this purpose.

References

Abbas, O.A.(2008). Comparisons between data clustering algorithms. *The International Arab Journal of Information Technology* **5**,320-5.

Aggarwal, C.C. and Reddy, C.K. (Eds).(2014). Data Clustering Algorithms and Application, CRC Press, USA.

Banfield, J. D. and Raftery A.E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics* **49** 803–821.

Ferligoj, A. and Batagelj, V. (1983). Some types of clustering with relational constraint. *Psychometrika* **48** 541-552.

Fraley, C. and Raftery, A.E. (2005). How many clusters? Which clustering method? Answers via model-based cluster analysis. *The Computer Journal* **41**, 578-588.

Goyal, V.K. (2014). An experimental analysis of clustering algorithms in data mining using Weka tool. *International Journal of Innovative Research in Science & Engineering* **2**, 171-6.

Han, J. and Kamber, M. (2006). Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers Inc, USA.

Jung, Y.G., Kang, M.S. and Heo, J. (2014). Clustering performance comparison using *K*- means and expectation maximization algorithm. *Biotechnol Biotechnol Equip* **28**, 44-8.

Kakkar, P. and Parashar, A. (2014). Comparison of different clustering algrithms using WEKA tool. *International Journal of Advanced Research in Technology, Engineering and Science* **1**, 20-2.

Kob, H.C. and Tan, G. (2005). Data mining applications in healthcare. Journal of Healthcare Information Management **19**, 64-72.

Schiopu, D., (2010). Applying TwoStep Cluster Analysis for Identifying Bank Customers' Profile. *Petroleum-Gas University of Ploiesti Romania*, **62**, 66- 75,

Sharma, N., Bajpai, A. and Litoriya, R. (2012). Comparison the various clustering algorithms of weka tools. *International Journal of Emerging Techonology and Advanced Engineering* **2**, 73-80.

SPSS Tecnical Report. (2001). The SPSS TwoStep Cluster Component, p.1-9.

Stein, I.L. (1935). Amenorrhea associated with bilateral polycystic ovaries. *Am J Obstet Gynecol* **29**,181.

Zhang, T., Raghu, R. and Miron, L. (1996). BIRCH: An efficient data clustering method for very large databases. In Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data. Canada, 4-6 July 1996.

Zheng, X., Cai, Z. Li, Q. (2005). An Experimental Comparison of Three Kinds of Clustering Algorithms. International Conference on Neural Networks and Brain Conference. China, 13- 15 October 2005.