COMPARISON OF RESULTS OF SELECTED CLUSTERING METHODS ON REAL DATA SET

Tomáš Löster

Abstract

The aim of this paper is to illustrate the possibilities and results of the process of clustering on the real data file (Wine fine) from *the UCI Machine learning Repository*. In current literature there are many methods and many distances measures, which can be mutually combined. In this paper were used different methods (Nearest neighbour, Farthest neighbour, Centroid method, Average distance, Ward's method) in combination with two selected distance measures (Euclidean and Mahalanobis). We compared situation with and without transformation of variables. There is no manual and rule which would clearly identify the appropriate combination method and distance measures during clustering. Simultaneously, in cluster analysis it is often necessary to determine the optimal number of clusters in to which the objects are to be classified. On the basis of the analyzes was found that in case of the Euclidean distance measure, it is preferable to transform the variables in using of the Ward method and the Farthest neighbour method. The success rates were higher than 80%. For other methods it was advisable to use methods without transformation of variables. The highest success in using of Mahalanobis' distance without transformation was again achieved with Ward's method. Using the Ward´s method in combination with the Euclidean distance measure we achieve a higher rate of success by 21.91% in comparison with Mahalanobis distance measure.

Key words: clustering, evaluating of clustering, methods, transormation

JEL Code: C 38, C 40

Introduction

Cluster analysis is multivariate method which objective is to classify the objects into groups called clusters. It is very often used statistical method, see e.g. (Halkidi et al., 2001; Löster at al., 2010; Řezanková et al., 2013; Žambochová, 2012). The need for creation of the groups of objects is an integral part of many disciplines. In practical tasks which are dealing with the

classification of objects is crucial for selecting the right multivariate classification methods if they are priory known or unknown the affiliations of the objects to clusters. Objects may be customers, patients, clients, documents, etc. Very often is used to classification of regions. Authors of papers very often used wages to describe regions. The problem of wages and poverty is described e.g. in (Bílková, 2011, 2012; Marek, 2013; Želinský, 2012). Other demographic variables, which are very often used in cluster analysis, are described in (Megyesiova, et al. 2011, 2012).

In the case when the investigated objects have known inclusion in the group, for classification is used the discriminant analysis, which aims to create a rule by which the new objects of unknown affiliation are classified. This is useful for example in medicine, where based on the properties of the patients the other patients are to be classified into groups known in advance. Second situation, i.e. when the classification of the objects is not known in advance is solved by cluster analysis. Currently there are many methods and approaches in current literature, which enable the analyst to classify number of objects set beforehand to clusters. Selection of possible combinations of methods is dependent on many factors.

Key role in cluster analysis play the similarity characteristics, resp. distances measures. Also in this case, the variable type, which characterizes each object, is critical. In case of quantitative variables the distance measures are used. There are many distance measures between objects. Linkage clustering methods and distance measures a whole series of combinations emerge, the choice is up to the analyst. Various combinations bring different results. In the current literature there are numbers of comparative studies that seek to evaluate various combinations of clustering methods and measure distances in a variety of conditions. However, there is not a clear rule that would strictly determine what combinations use in what situations. Although they are indicated for instance situations in which different distance measures are unsuitable (for example in case of a strong correlation between the input variables), but the actual effect of breaking of this assumption is usually not analyzed. In the same way the advantages and disadvantages of different clustering algorithms are indicated.

The aim of the paper is to show results of clustering on real data file – Wine file from *the UCI Machine learning Repository*.

1 Clustering methods

The aim of cluster analysis is the classification of objects, see (Gan et al., 2007). There are various methods and procedures to do that. These methods and procedures can be

categorized according to various criteria see e.g. (Gan et al., 2007; Řezanková et al., 2009). Mostly they are divided on traditional methods and new approaches in the literature. Traditional methods are well developed and they are applied in many software products.

In current literature there are numbers of clustering algorithms, which are implemented to many specialized software products. Application of various methods of clustering on same objects described by identical properties can produce different results. As stated by Gan et al. (2007) and Halkidi (2002) "It cannot be a priori said which method is the best for a given problem. Usually, the method of the nearest neighbour is the least suitable and method of average distance or Ward's method suits in many cases the best". But it is important also those practical experience researchers with the type of job are used. Among the methods hierarchical clustering can be included, for example, the nearest neighbour method, method of the farthest neighbour, method of the average distance, centroid method.

Nearest neighbour method it is the oldest and the simplest method. There are searched two objects, between which the distance is the shortest and they are joined to the cluster. Another cluster is created by linking the third closest object. Distance between two clusters is defined as the shortest distance of any point in cluster in relation to any point in another cluster, see Gan et al. (2007). As one of crucial disadvantage of this methods is stated that occurs so-called *chaining*, when two objects, which are the closest in relation to each other, but not in relation to majority of other objects, are sorted to one cluster.

Farthest neighbour method is based on the opposite principle than the method of the nearest neighbour. The advantage of this method is that it creates small, compact and clearly separated clusters. Contrary to the nearest neighbour method there is no problem with clusters' chaining.

Using **method of average distance** the criterion for emerge of the clusters represents the average distance of all objects in one cluster to all objects in second cluster. Results of this method are not influenced by extreme values as in the case of method of the nearest and furthers neighbour. Emerge of the cluster is dependent on all objects.

Centroid method was firstly used by Sokal and Michener under name "weighted group method".This method does not use between-cluster distances of the objects. To new cluster those two clusters are merged, between what is minimal distance of their centroids, while the

centroid is understood as an average of the variables in particular clusters. The advantage of this method is that it is not that significantly influenced by remote objects.

Median method was firstly introduced by Gower under name "unweighted group method". The aim of the method is the effort to eliminate the disadvantages of centroid methods, see above. Gower proclaimed that "... different number of objects of clusters cause different weight of first two parts of the recursive prescription of centroid method and thus it happens that the characteristics of small clusters disappears in final linkage". Median method is an analogy of centroid method and the difference is that instead of the distance between centroid clusters is used the distance between medians of those clusters. To one cluster are merged two clusters between which medians is the closest distance. The advantage of this method is in removing of different weights which are in centroid method assigned to differently sized clusters.

Ward's method solves the clustering procedure differently than above stated methods that are optimizing the distances between particular clusters. Method minimizes the heterogeneity of clusters, i.e. clusters are formed using maximization of intragroup homogeneity. As the measure of homogeneity of clusters is understood intragroup sum of squares of the deviations of values from the average of the clusters and it is called *Ward's criterion*. Criterion for linking the clusters is based on the idea that in each step of clustering there is minimal increment of Ward's criterion. Ward's method has tendency to remove small clusters and create clusters of approximately same size.

Besides the clustering methods themselves and important (key) role is played also by the measures of dissimilarity. Similarity is used as the criterion for the creation of clusters. Measurement of the similarity of objects when they are characterized by quantitative variables is based on the distances of the objects. Transformation of the distance measures to similarity (dissimilarity) measures is done according to simple rules. Very important are the measures of similarities, resp. the distance levels. There are a number of distance levels and in the practice they are combined with various clustering methods, see e.g. (Gan et al., 2007; Řezanková et al. 2009).

For measurement of the distance are frequently used:

Euclidean distance represent the length of hypotenuse of a rectangular triangle. Calculation of this measure is based on Pythagoras theorem. **Mahalanobis distance** diminishes the problem while using non-standardized data that can cause differences among clusters due to different measurement units. This measure is usable in the case when all the variables characterizing the objects are mutually correlated.

Detailed descriptions of methods and formulas of particular distance measures can be found e.g. in Řezanková (2009) or Gan et al. (2007).

2 Wine file

The wine file data set contain informations, which are the results of a chemical analysis of wines grown in the same region in Italy but derived from three different cultivars. The analysis determined the quantities of 13 constituents found in each of the three types of wines. Total number of objects is 178 samples of wine.

The attributes are: Alcohol, Malic acid, Ash, Alcalinity of ash, Magnesium, Total phenols, Flavanoids, Nonflavanoid phenols, Proanthocyanins, Color intensity, Hue, OD280/OD315 of diluted wines, Proline. All attributes are continuous.

For Classification 178, five of the above-mentioned clustering methods were used. They were combined with the Euclidean and Mahalanobis distance measure. At the Euclidean distance measure, the results were compared with the clustering without and with the transformation of the variables.

Tables 1 – 3 show the results of clusters obtained on the basis of the Ward method in combination with the Euclidean distance measure and the Mahalanobis distance measure.

Table 1 shows the results of clustering using the Euclidean distance measure. No transformation of variables was applied when cluster methods were applied.

Tab. 1: Results of Clustering (Ward´s method, Euclidean distance, without transformation)

Variety/Cluster	1	2	3	Total
1	46	0	13	59
2	2	51	18	71
3	0	21	27	48
Total	48	72	58	178

Source: our calculations

Table 1 shows that 46 samples of wines from the total of 59 samples of the first variety were correctly classified. For the second variety, 51 samples were classified correctly. For the third variety, 27 samples were correctly classified.

Table 2 shows the results of clustering once again using the Euclidean distance measure. In this application, transformation of variables was applied.

Tab. 2: Results of Clustering (Ward´s method, Euclidean distance, with transformation)

Variety/Cluster	1	2	3	Total
1	59	0	0	59
2	5	58	8	71
3	0	0	48	48
Total	64	58	56	178

Source: our calculations

Table 2 shows that all 59 samples of wines of the first variety were correctly classified. For the second variety, 58 samples of wines were classified correctly. For the third variety, all samples were correctly classified.

Table 3 lists the clustering results for the Mahalanobis distance measure. In this application, transformation of variables was not applied.

Tab. 3: Results of Clustering (Ward´s method, Mahalanobis distance, withnout transformation)

Variety/Cluster	1	2	3	Total
1	27	32		59
2		71		71
3		25	23	48
Total	27	128	23	178

Source: our calculations

Table 3 shows that 27 samples of the first variety were correctly classified. For the second variety, all 71 samples of wines were correctly classified. For the third variety, 23 samples were correctly classified out of a total of 48.

Table 4 shows the comparison of the results of the success of classification of individual clustering methods using the Euclidean distance measure. Transformation procedures are compared and if the transformation was not used.

Tab. 4: Comparison of results for the Euclidean distance measure

Methods	With transformation	Without transformation	difference
Nearest neighbour	37,64%	42,70%	**5,06%**
Farthest neighbour	83,71%	67,42%	**16,29%**
Centroid method	37,64%	61,24%	**23,60%**
Average distance	38,76%	53,93%	**15,17%**
Ward's method	92,70%	69,66%	**23,03%**

Source: our calculations

Table 4 shows that the best result was achieved when the Ward method was used and transformation applied to the original variables. In this case, 92, 7% of objects were correctly classified. The greatest difference in use and non-use of the transformation was achieved with the Centroid method. If the transformation was not applied, the classification was about 23, 6 % higher.

Table 5 shows a comparison of the success of the classification of the individual clustering methods using the Mahalanobis distance measure. Processes without transformation are compared. The table again shows that the best result was achieved using the Ward method - the success rate of 70, 79%.

Tab. 5: Comparing the results of the Mahalanobis distance

Methods	Without transformation
Nearest neighbour	38,76%
Farthest neighbour	35,96%
Centroid method	38,76%
Average distance	38,76%
Ward's method	70,79%

Source: our calculations

Table 6 shows a comparison of the results of the success of classification of individual clustering methods using Euclidean and Mahalanobis distance measurements in a situation where no transformation was applied.

Tab. 6: Comparison of the results the Euclidean and Mahalanobis distance measure (without transformation)

Methods/Distance	Euclidean	Mahalanobis	difference
Nearest neighbour	42,70%	38,76%	**-3,93%**
Farthest neighbour	67,42%	35,96%	**31,46%**
Centroid method	61,24%	38,76%	**-22,47%**
Average distance	53,93%	38,76%	**-15,17%**
Ward's method	69,66%	70,79%	**-1,12%**

Source: our calculations

Table 6 shows that if the transformation is not used, it is preferable to use the Mahalanobis distance measure, except in the case of the most distant neighbor method

Conclusion

Cluster analysis is a multivariate statistical method, which is used to classify objects into clusters. There are many clustering methods and there are many measures of the distances between objects. The combination of various method and different distance measures give different results. The current literature does not address the different combinations and there is no indication which combination is successful.

This article uses the Wine file from database *UCI Machine learning Repository*. The wine file data set contain informations, which are the results of a chemical analysis of wines grown in the same region in Italy but derived from three different cultivars. The analysis determined the quantities of 13 constituents found in each of the three types of wines.

In order to compare the clustering results of 178 samples of wines, a total of five aggregation methods and two measurements of distance were used. The methods of clustering have been chosen: Nearest neighbour, Farthest neighbour, Centroid method, Average distance, Ward's method. These methods differ in the time of occurrence and the way of clustering. Each of these methods was used both in combination with the Euclidean distance measure and the Mahalanobis distance measure. In addition, the Euclidean distance measure have been compared to its use without and with transformation. On the basis of the analyzes performed, it was found that in the case of the Euclidean distance measure, it is more appropriate to transform the variables in the Ward method and the Farthest neighbour. Their success rate was higher than 80%. For other methods it was advisable to use methods without transformation of variables. The highest success in using Mahalanobis' distance without transformation was again achieved with Ward's method. Using the Ward´s method in combination with the Euclidean distance measure we achieve a higher success rate by 21.91% (in comparison with using of Mahalanobis distance measure).

Acknowledgment

This paper was supported by long term institutional support of research activities IP400040 by Faculty of Informatics and Statistics, University of Economics, Prague, Czech Republic.

References

Bílková, D. (2011). Modelling of income and wage distribution using the method of l-moments of parameter estimation. In Löster Tomas, Pavelka Tomas (Eds.), International Days of Statistics and Economics (pp. 40-50). ISBN 978-80-86175-77-5.

Bílková, D. (2012). Development of wage distribution of the Czech Republic in recent years by highest education attainment and forecasts for 2011 and 2012. In Löster T., Pavelka T. (Eds.), 6th International Days of Statistics and Economics (pp. 162-182). ISBN 978-80-86175-86-7.

Gan, G., Ma, Ch., Wu, J. (2007). *Data Clustering Theory, Algorithms, and Applications*, ASA, Philadelphia.

Halkidi, M., Vazirgiannis, M. (2001). *Clustering validity assessment: Finding the optimal partitioning of a data set*, Proceedings of the IEEE international conference on data mining, pp. 187-194.

Löster, T., & Langhamrova, J. (2011). Analysis of long-term unemployment in the Czech Republic. In Löster Tomas, Pavelka Tomas (Eds.), International Days of Statistics and Economics (pp. 307-316). ISBN 978-80-86175-77-5.

Marek, L. (2013). Some Aspects of Average Wage Evolution in the Czech Republic. In: International Days of Statistics and Economics. [online], Slaný: Melandrium, pp. 947–958. ISBN 978-80-86175-87-4. URL: http://msed.vse.cz/files/2013/208-Marek-Lubos-paper.pdf.

Megyesiova, S., & Lieskovska, V. (2011). Recent population change in Europe. In Löster Tomas, Pavelka Tomas (Eds.), International Days of Statistics and Economics (pp. 381-389). ISBN 978-80-86175-77-5.

Megyesiova, S., & Lieskovska, V. (2012). Are europeans living longer and healthier lives?. In Löster Tomas, Pavelka Tomas (Eds.), 6th International Days of Statistics and Economics (pp. 766-775). ISBN 978-80-86175-86-7.

Meloun, M., Militký, J., Hill, M. (2005): Počítačová analýza vícerozměrných dat v příkladech, Academia, Praha.

Řezanková, H., Húsek, D., Snášel, V. (2009). *Cluster analysis dat,* 2. vydání, Professional Publishing, Praha.

Řezanková, H., & Löster, T. (2013). Shlukova analyza domacnosti charakterizovanych kategorialnimi ukazateli. *E+M. Ekonomie a Management, 16*(3), 139-147. ISSN: 1212-3609.

Šimpach, O. (2012). Statistical view of the current situation of beekeeping in the Czech Republic. In Löster Tomas, Pavelka Tomas (Eds.), 6th International Days of Statistics and Economics (pp. 1054-1062). ISBN 978-80-86175-86-7.

Stankovičová, I., Vojtková, M. (2007): Viacrozmerné štatistické metódy s aplikáciami, Ekonómia, Bratislava.

Žambochová, M. (2012): *Classification in terms of students' preferences for information sources, Efficiency and responsibility in education.* 9th International Conference on Efficiency and Responsibility in Education, Praha, p. 612-620, ISBN 978-80-213-2289-9.

Želinský, T., & Stankovičová, I. (2012). Spatial aspects of poverty in Slovakia. In Löster Tomas, Pavelka Tomas (Eds.), 6th International Days of Statistics and Economics (pp. 1228-1235). ISBN 978-80-86175-86-7.

Contact

Ing. Tomáš Löster, Ph.D.
University of Economics, Prague,
Dept. of Statistics and Probability
W. Churchill sq. 4,
130 67 Prague 3, Czech Republic
losterto@vse.cz