DOI: 10.5604/20830157.1201309

MODIFIED ALTERNATIVE DECISION RULE IN THE PRE-CLUSTERING ALGORITHM

Volodymyr Mosorov, Taras Panskyi, Sebastian Biedron
Lodz University of Technology, Institute of Applied Computer Science

Abstract. *In this paper the pre-clustering algorithm with the modified decision rule has been presented. The application of pre-clustering algorithm answers the question whether to carry out the clustering or would it result in the appearance of artificial structure (input data is one cluster and it is unnecessary to divide it). The versatility and simplicity of this algorithm allows using it in a various fields of science and technology. The pros and cons of pre-clustering algorithm have been also considered.*

Keywords: pre-clustering, pre-cluster, decision rule

ZMODYFIKOWANA ALTERNATYWNA REGUŁA DECYZYJNA W ALGORYTMIE WSTĘPNEGO KLASTROWANIA

Streszczenie. *W tej pracy został przedstawiony algorytm wstępnego klastrowania oraz zmodyfikowana alternatywna reguła decyzyjna. Zastosowania algorytmu wstępnego klastrowania odpowiada na pytanie czy potrzebna procedura klastrowania czy spowodowałoby to pojawienia sztucznej struktury (dane wejściowe są jednym klasterem i nie ma potrzeby podziału). Uniwersalność i prostota tego algorytmu pozwala na wykorzystanie go w różnych dziedzinach nauki i techniki. Zalety i wady algorytmu wstępnego klastrowania zostały również rozważone.*

Słowa kluczowe: wstępna klasteryzacja, reguła decyzyjna

Introduction

The concept and application of clustering is quite wide, together with well known clustering algorithms and approaches it is repeatedly described [4].Therefore, it is reasonable to avoid the known details of cluster analysis, its application in various fields of science and technology, and popular clustering algorithms [1, 10], and focus on the proposed [9] pre-clustering algorithm.

Pre-clustering is the procedure of checking the possibility of clustering the input data. Checking this possibility answers the question whether data can be divided into more than one cluster. Pre-clustering algorithm [9] unlike existing does not require a priori information about the location of clusters and additional control tools, such as thresholds, measures of the object similarity. The pre-clustering algorithm is a universal and perspective algorithm for the input data primary analysis.

The universality of pre-clustering algorithm is explained by its ability to use all kinds of numeric attributes. A numeric attribute is quantitative; that is, it is a measurable quantity, represented in integer or real values. On the other hand the universality is achieved by the possibility of applying this algorithm to the most continuous distribution laws (normal distribution, Student's distribution, Weibull distribution etc.).

Despite the advantages of the pre-clustering algorithm and simplicity of its decision rule it has several disadvantages that cause imposing some limitations for the proper finding the number of clusters. The main disadvantage of this rule is the dependence of the results from the computed average distances. If clusters include anomalies (rare objects that are located at a great distance in relation to other objects in the cluster) the average distance calculation result is highly dependent on these objects. This drawback strongly influences on the decision rule especially when the input data are not infinitely large and include a limited number of objects. To eliminate the strong influence of isolated objects on a decision rule the modification of the existing rule has been proposed.

The main objective of the analysis of the input data is the answer the question whether to carry out clustering, or input data have no internal structure and clustering process will not result in its discovery but only to artifact appearance (artificial structure).

1. Analysis of published data and problem statement

One of the most known pre-clustering algorithms require a user setting of certain input parameters, one of the examples is a canopy clustering algorithm, presented by [8]. It is often used for the preliminary analysis of input data or for primary clusterization for the k–means algorithm or hierarchical clustering algorithm. The disadvantage of this pre-clustering algorithm is the heuristic definition of two threshold values (distances) T1 and T2.

Another example is a BIRCH pre-clustering algorithm [6]. This algorithm is an efficient data reduction method in the case of large data sets. However, BIRCH requires the set of the optimization key parameters (like branching factor, quality threshold and selection of the separator line).

Some clustering algorithms are part of already created algorithms and make up its preprocessing step [5]. For example, an algorithm for preprocessing k–means clustering for better and more effective number of clusters determination.

Clustering refers to unsupervised learning and, for that reason it has no a priori data information (distribution, the number of possible clusters, labeled attributes). However, to get good results, the clustering algorithm depends on input parameter (for ex. number of clusters). In this context the question what number of clusters is optimal comes into existence. Today the usage of validation criteria is the possible options of answering the posed question. Validity indexes allow us to estimate and select the optimum value of the input clustering parameter. Validity indexes are divided into internal and external. External validation is based on previous knowledge about data. The main external criteria's [2, 3]: Rand index, Jaccard index, Fowlkes–Mallows index. In the real the world in the practical problems is not always possible receive input data priori information. Therefore, external criteria rarely find their application in the primary analysis of input data clustering. Internal indexes are based on information within the cluster. The main internal criteria's: Davies-Bouldin index, cluster density, average within centroid distance [7]. Internal validation criteria do not require a priori information about input data. However, the usage of one criterion will not cause absolute reliability of clustering results. So, it is advisable to use as more criteria as possible for increasing the reliability of the clustering results. The majority of validation criteria are based on a multiple choice and the substitution of input clustering parameters (for example, the number of the clusters) and on the choice of the most optimal input parameter. One of the disadvantages is the strong user dependence, since even if the criteria for result validation are used, the input parameters are likely to be chosen erroneously.

The idea of pre-clustering algorithm is to eliminate the user influence on the clustering results. Pre-clustering algorithm offers the possibility of "artificial intelligence" that is without aprioristic input data information, and without additional control tools (eg multiple testing, selection of optimum likelihood criterion) to determine whether the number of cluster could be more than one.

The main task of input data analysis is an answer to the question whether it is necessary to perform data clustering or input data have no inner structure and the clustering process will result not in its revealing but in occurrence of artifacts (artificial structures). The pre-clustering algorithm allows us to analyze and evaluate input data and decide whether input data represent a single cluster which does not require further clustering, or two independent separate clusters which can be identified as a further clustering opportunity.

2. Pre-clustering algorithm

Pre-clustering is a procedure of checking the possibility of input data clustering. The pre-clustering algorithm forces the division of input data set into two pre-clusters. The pre-cluster is a group of objects which is not a single cluster, but can become one after checking. To decide whether a given pre-cluster is a single cluster or a part of a bigger cluster, the pre-clustering algorithm has been used.

After the forced division of the input data set, the heuristic decision rule of the pre clustering algorithm uses the estimated average distances between objects in the found pre clusters $d(K_1)$ and $d(K_2)$, pre-clusters K_1 and K_2 accordingly and average distances between objects $d(K)$ of the whole input data set K using the Euclidean distance in the 2D space. The decision rule estimates the possibility of the pre-cluster to be a cluster. In pre-clustering algorithm experimental estimations 2D-data sets are used in order that the reader is able to visually verify the validity of the results (i.e., how well the clustering algorithm discovered the clusters of the data set). In the case of large multidimensional data sets (e.g. more than three dimensions) effective visualization of the data set would be difficult. Moreover the perception of clusters using available visualization tools is a difficult task for humans that are not accustomed to higher dimensional spaces. However, the pre-clustering algorithm is flexible enough for analyzing multidimensional data. In the case of multidimensional data, the parameters of the decision rules are logically modified (e.g. Euclidean distance is calculated taking into account the number of attributes).

The pre-clustering algorithm as opposed to other existed algorithms does not require a priori information about cluster location, distribution of objects and about additional means of control (as, for example, threshold meanings or measures of object similarity) for proper detecting whether the number of clusters is larger then one. This preclustering algorithm is multipurpose and promising for a primary analysis of investigated input data.

The universality of the pre-clustering algorithm can be explained by the ability of its using to all kinds of numerical attributes, that is, measured numerical quantities produced as integral or real values. On the other hand, the universality is achieved by the possibility of applying this algorithm to the majority of continuous distribution laws (normal distribution, truncated normal distribution, Student's t–distribution, uniform distribution, Weibull distribution and others).

In spite of advantages of the pre-clustering algorithm and the simplicity of its decision rule has some sufficiently serious disadvantages which cause the introduction of some limitations for the proper algorithms action. The main disadvantage of the decision rule is the dependence of the results on the calculated average distances. If clusters include isolated objects or anomalies (single objects located at a large distance from other cluster objects), the results of calculating the average distances become strongly dependent on these objects which results in wrong decision making. This disadvantage strongly influences on the decision making particularly in cases when input data set is not infinitely large and includes the limited number of objects (for example, less than 100).

For eliminating the strong influence of isolated objects on decision making the modification of the existed decision rule is proposed.

3. Modified decision rule

The idea of pre-clustering algorithm represented in the form block scheme in Fig. 1.

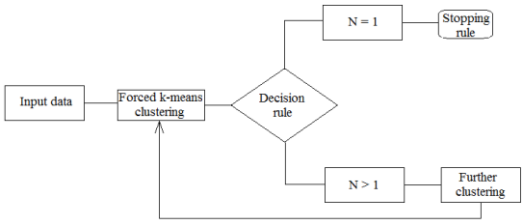


Fig. 1. Block scheme of a pre-clustering algorithm

The disadvantage of decision rule (as well as in all the clustering partitioning algorithms) is a strong dependence of computed distances on the nature of input data. If the input data are globular or spherical form (presume the symmetric Gaussian form), with the possible existence of mutually exclusive clusters, in this case the decision rule will work properly. However, when the input data make arbitrary form, include anomalies, the decision rule could work incorrectly. To reduce the effect of the factors described above on the decision rule, the replacement of calculation of the input data's average distances to the average distances from the pre-cluster's density center to all the objects in the selected pre-cluster has been proposed.

The onset of action of the pre-clustering algorithm is unchanged, that is the compulsory k-means (c–means, k–medians or other clustering algorithm that requires setting the number of clusters as an input clustering parameter) input data clustering is carried out. After forced division every pre-cluster is checked by the decision rule, which subsequently determine whether this pre-cluster is a separate cluster, or a part of a larger cluster. In this meaning pre-cluster is group of objects formed after the forced division. Pre-cluster is not always a cluster, but it could become a cluster if the decision rule signalize about this.

Modified pre-clustering algorithm and its decision rule requires the following steps:

1) A forced k-means clustering or other clustering which requires setting the number of clusters as an input parameter is carried out. In this algorithm forced clustering always divides analyzed input data into strictly two pre-clusters.
2) In each divided pre-cluster the average distance from the center to all objects within a pre-cluster is calculated.
3) The average distance between all objects before the forced division is calculated, that is the input intra data distance $d(K)$, where K is a whole input data set.
4) The possibility of cluster existence is checked by the decision rule, i.e. the possibility of divided pre-clusters to be clusters.

Modified decision rule in a pre-clustering algorithm could be written in this form:

$$decision = \begin{cases} N=1, & R(K_1)+R(K_2) > d(K) \quad and \\ & C_1((x_1,y_1)), R(K_1)) \cap C_2((x_2,y_2)), R(K_2) \neq \varnothing, \\ N>1, & otherwise, \end{cases} \tag{1}$$

where $C(z,r) = \{s \in \Re^2 : \|s-z\| \leq r\}$.

After the forced clustering for each found pre-cluster K_1 and K_2 the average distance from the pre-cluster's center to all objects inside it $R(K_1)$ and $R(K_2)$ is calculated. Then the circle whose center coincides with the pre-cluster center is built, (x_1, y_1) is the center of the first pre-cluster and (x_2, y_2) is the center of the second pre-cluster. If the first inequality of the decision rule is satisfied and built circles intersect (circle C_1 intersects circle C_2), it shows that in this data set one cluster exists (pre-clusters divided by forced clustering are not separate clusters). In all other cases found pre-clusters are independent ones.

4. Choosing the center of the pre-cluster

In the modified decision rule the average distances from all objects of the pre-cluster to its center are calculated. In many popular algorithms the center of the group of objects is denoted in different ways. In the k–means algorithm the center of the group is considered to be a centroid. The centroid is a mean value of all analyzed objects in one group. In the k–medians algorithm the median of the group of objects is calculated instead of calculating the mean value of all objects in the group for determining the centroid.

The proposed decision rule determines the center of the group as a local density maximum of the group of objects (before clustering) or of the pre-cluster (after clustering). The most significant disadvantage of choosing group centers (centroids, or medians) is its strong dependence on anomalies.

At the high object's density big data sets, in the case of a group's globular form, the difference between the centroid, median and maximum density center is insignificant, however, when the objects form the group of arbitrary shape with the variable density, the difference between the centroid, median and maximum density center becomes bigger.

Choosing the maximum density center of the group of objects is explained by the fact that the input data set (without a priori information about it) can contain any number of anomalies. The centroid and median are influenced by this factor and can react inadequately, which can cause erratic results, but the density of the group of objects is resistant to anomalies and their influence on it.

Presented modified decision rule is similar to the known the criterion of spherical resolution, when the sum of radii of two groups of objects is less than the distance between their centers (in such a data set only one cluster exists). In the criterion of spherical resolution the center of the cluster is a centroid and its radius is determined as maximum distance from the center of the cluster, or the radius of the least circle surrounding all objects in the cluster.

The disadvantage of the criterion of spherical resolution is the fact that maximum radius from the center of the group of objects heavily depends on the anomalies. At the significant standard deviation and in the presence of anomalies the criterion of spherical resolution causes the distortion of the results.

5. Experimental results obtained by the modified decision rule

The versatility of presented algorithm is based on its application to various practical problems. For example, experimental data are images with defects: human skin (application in medicine), materials (industrial application), and are presented at the table 1.

The input image of size 256×256 pixels is first divided into rectangular regions, each of which is of size 16×16 pixels. The size of the region determines detection accuracy and can be empirically chosen. For each region of the image the mean and standard deviation have been found. Thus the image has been transformed into a data set. Then the forced k-means clustering have been carried out (k = 2). Using the decision rule have been defined whether an input data set is more than one cluster. Image without defects is considered as one cluster.

Most of the popular algorithms for image analysis easily detect the presence or the number of defects. However, the experimental images are presented only for visual comparison of the veracity of decision rule. In practical problems there is no prior information about the number of clusters, ie, *there is no input images* and the data are presented as a set of objects.

The first case shows that the conditions of the decision rules are satisfied, and as a result one cluster is detected in the input data set.

At the second case both condition of the decision rule are satisfied. In given data set two separate clusters exist, but it is still possible that it makes up one general cluster. In this case it is necessary to use additional means of checking and control (tests, criteria).

As illustrated in the third case, forcibly divided data set formed two separate pre-clusters.

The fourth case demonstrates that pre-clusters are located in the significant distant from each other. As a result of analysis, two independent clusters are located in the input data set.

The analysis of the fifth case demonstrated the existence of two separate clusters. On this image the colour and structure of skin look like normal and using the decision rule can cause inadequate results. Such a set of input data can be analyzed as one elongated cluster, though actually there are two of them. In such case one more parameter should be added to the standard deviation and mean value and n-D analysis should be performed.

Table 1. Experimental results

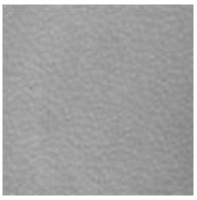
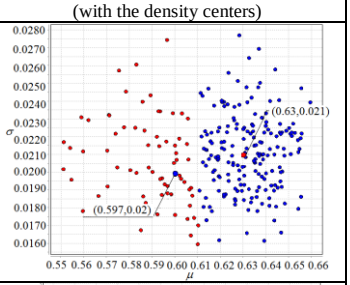
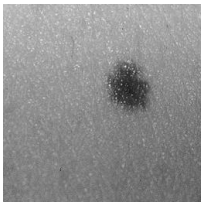
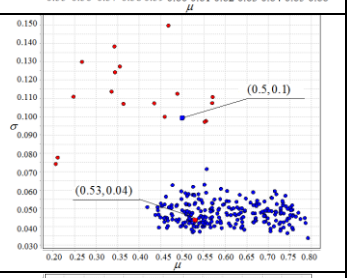
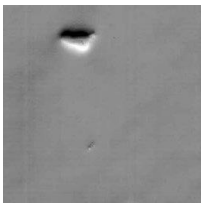
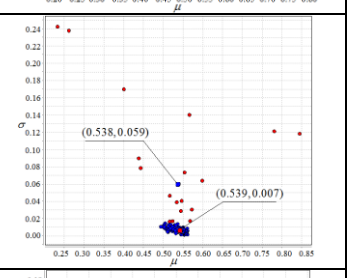
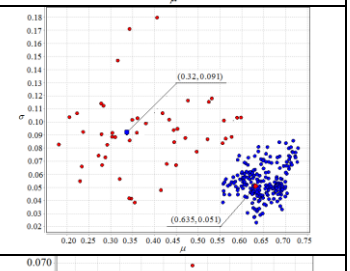
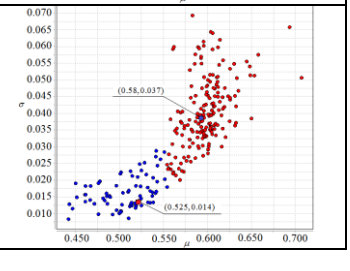
No.	Testing image 256×256	Visualization of forced k-means clustering in the 2D attribute space (with the density centers)
1		
2		
3		
4		
5		

Table 2 shows the comparison of the results obtained by the decision rule and the criterion of spherical resolution at the condition that normal skin image is considered to be one single cluster.

Table 2. Experimental results

No.	The number of clusters (visual analysis)	The number of clusters (modified decision rule)	The number of clusters (criterion of spherical resolution)
1	1	1	1
2	2	2 (additional means of checking)	1
3	2	2	1
4	2	2	1
5	2	2 (additional parameter)	1

The decision rule is not always precise, but for primary analysis of clustering possibilities the use of this rule together with the pre-clustering algorithm provides the stronger probability of correct cluster detecting than in the case of criterion of spherical resolution.

Conclusions

In this article the modified decision rule in the pre-clustering algorithm has been presented. Also this decision rule was tested on a series of experimental data, and the results were compared with the criteria of spherical resolution. The modified decision rule allows to obtain a better results than classical, one and much better than criterion of spherical resolution.

In this article experimental data were divided into one or two clusters, but if the input data contain more than two clusters the stopping rule for the pre-clustering algorithm should be applied.

This decision rule has its disadvantages. One of them is still the dependence of the parameters on calculated distances. When objects are significantly scattered and their number is small, there are possibilities for existing anomalies and, accordingly, the difficulties in obtaining adequate results.

In further investigations it is proposed not to calculate distances and pay attention only to the density of the objects.

References

[1] Aggarwal C.: Data Clustering: Algorithms and Applications 1st Edition, Chapman and Hall, 2013.
[2] Gan G., Ma C., Wu G.: Data Clustering: Theory, Algorithms and Applications. ASA-SIAM Series on Statistics and Applied Probability, 2007.
[3] Hofmann M., Klinkenberg R.: RapidMiner: Data Mining Use Cases and Business Analytics Applications, Chapman and Hall/CRC, 2013.
[4] Jain A., Murthy M.: Flynn P.: Data Clustering: A Review. ACM Computing Surveys (CSUR), 1999.
[5] Khan M.A.: H Pre-processing for K-means Clustering Algorithm. Senior Projects Spring, 2015.
[6] Kovács L., Bednarik L.: Parameter Optimization for BIRCH Pre-Clustering Algorithm. 12th IEEE International Symposium on Computational Intelligence and Informatics, 2011.
[7] Liu Y., Li Zh., Xiong H., Gao X., Wu J.: Understanding of Internal Clustering Validation Measures, IEEE International Conference on Data Mining, 2010.
[8] McCallum A., Nigam K., Ungar L.H.: Efficient Clustering of High Dimensional Data Sets with Application to Reference Matching, Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining, 169–178.
[9] Mosorov V., Tomczak L.: Image Texture Defect Detection Method Using Fuzzy C–Means Clustering for Visual Inspection Systems, Arabian Journal for Science and Engineering, 2014.
[10] Rokach L., Maimon.: Clustering Methods, Data Mining and Knowledge Discovery Handbook, 2005.

D.Sc. Eng. Volodymyr Mosorov
e-mail: w.mosorow@kis.p.lodz.pl

Volodymyr Mosorov received his Ph.D. in 1998 from the State University of Lviv, Ukraine. V.Mosorov was awarded the title of Doctor of Science from AGH University of Science and Technology Krakow Poland in 2009. He is now an associate professor at the Institute of Applied Computer Science of Lodz University of Technology, Poland. His research interests include data mining and clustering. He has been involved in these areas for more than 15 years. Member of the The *Polish* Information Processing Society. He has published more than 80 technical articles.

M.Sc. Taras Panskyi
e-mail: tpanski@kis.p.lodz.pl

Graduate of the Electrotechnics Department at the Lviv National Polytechnic University, Ukraine. From 2013, Ph.D. student at the Institute of Applied Computer Science of Lodz University of Technology, Poland. His research interests include data clustering, reliability and availability indexes of embedded systems, educational migration.

M.Sc. Sebastian Biedron
e-mail: sbiedron@wpia.uni.lodz.pl

A graduate of the Department of Science and Mathematics at Lodz University. From 2012 year is a court expert at the District Court at the Prague. From 2013, Ph.D. student at the Institute of Applied Computer Science of Lodz University of Technology. Supervisor of the Ph.D. thesis is D.Sc. Eng. Volodymyr Mosorov, prof. Lodz University of Technology.

otrzymano/received: 10.10.2015 *przyjęto do druku/accepted: 29.02.2016*