18. Золотарев, В. М. Устойчивые законы и их применение [Текст]: монография / В. М. Золотарев. – Москва, 1984. – 66 с.
19. Леви, П. Стохастические процессы и броуновское движении [Текст]: монография / П. Леви. – М.: Наука, 1972. – 375 с.
20. Стрелкова, Т. А. Статистические свойства выходных сигналов оптико-телевизионных систем с ограниченным динамическим диапазоном [Текст] / Т. А. Стрелкова // Восточно-Европейский журнал передовых технологий. – 2014. – Т. 2, № 9 (68). – С. 38–44. doi: 10.15587/1729-4061.2014.23361
21. Strelkova, T. A. Studies on the Optical Fluxes Attenuation Process in Optical-electronic Systems [Text] / T. A. Strelkova, // Semiconductor physics, quantum electronics & optoelectronics (SPQEO). – 2014. – Vol. 17, Issue 4. – P. 421–424.

В даній статті представлений спосіб автоматизованої оцінки числа кластерів, заснований на евристичному підході чіткої кластеризації вхідного масиву даних з використанням густини розподілу даного масиву. Спосіб включає в себе правило прийняття рішення. Даний підхід чіткої кластеризації дає змогу оцінити, чи є число кластерів більше за одиницю

Ключові слова: спосіб кластеризації даних, кластер, евристичний підхід, густина розподілу

В данной статье представлен способ автоматизированной оценки числа кластеров, основанный на эвристическом подходе четкой кластеризации входного массива данных с использованием плотности распределения данного массива. Способ включает в себя правило принятия решения. Данный подход четкой кластеризации позволяет оценить, является ли число кластеров больше чем единица

Ключевые слова: способ кластеризации данных, кластер, эвристический подход, плотность распределения

UDC 004.9

DOI: 10.15587/1729-4061.2015.39785

NEW DATA CLUSTERING HEURISTIC ALGORITHM

V. Mosorov
Doctor of Technical Sciences*
E-mail: volodymyr.mosorov@p.lodz.pl

T. Panskyi
Graduate student*
E-mail: panskyy@gmail.com

*Institute of Applied Computer Science
Lodz University of Technology
Stefanowskiego str., 18/22,
Lodz, Poland, 90-924

1. Introduction

The goal of this article is to propose a method of clustering the data by data mining. Clustering is a division of data into groups of similar objects using appropriate algorithms and rules. Clustering has been thoroughly studied and perfected over the years in different areas and fields of science including pattern recognition [1], machine learning [2], statistics and image processing [3]. Data clustering has been applied in different fields of science and industry, such as biology and bioinformatics, medicine, marketing, computer science, social science, robotics, mathematical chemistry, climatology, physical geography and others.

Traditional clustering methods and algorithms are computationally expensive when clustering is applied to large data sets. There are three possible cases when the data set can be determined as large: when in one data set there are a lot of elements, when one element has many special properties and when the correct definition of a large number of clusters is problematic. In our case the large data set is the set with a lot of elements [4]. Considering the development of computer technology, consumption of time and computer memory to perform clustering is not taken into account.

2. Analysis of published data and problem statement

Application of various clustering algorithms is impressive: from the clustering of textual data, images, videos to a network clustering algorithm along with the real-time streaming data and outlier detection. Annually old algorithms are modernized and improved both a new ways and approaches of clustering techniques are investigated. Clustering algorithms can be broadly classified [5] into three categories: partitioning, hierarchical and density based. The proposed algorithm has found his niche in density based algorithms as a basic input parameter is the density distribution of a primary data. This algorithm has certain advantages in comparison with existing density based algorithms that will be discussed after viewing the publications. In comparative analysis, attention was paid to some representatives of density based clustering algorithms: DBSCAN [6], FDBSCAN [7], ODBSCAN [8], VDBSCAN [9], ST-DBSCAN [10], Incremental DBSCAN [11], and RDBC [12]. Despite the advantages such as: an opportunity to find arbitrary shape cluster, remove noise from the dataset, cluster spatial-temporal data according to non-spatial, these algorithms has the main drawback – determining the initial parameters for the correct application of the algorithm, such as radius, minpts, number of identical circles. Developed algorithm does not use any of the above input parameters only the initial parameter is the input data set.

3. Purpose and objectives of the study

The key purpose of this paper is determining the quantity of clusters, namely, we are interested in whether there is more than one cluster using only a data set as an input parameter.

In accordance with the set goal the following research objectives are identified:

1. Development a decision rule for clustering algorithm for appropriate estimation the number of clusters in data set.

2. Preliminary assessment of input data which includes determining the height and number of local maxima, calculation the number of points that could fall in each region when changing density distribution of data set.

4. Overview of clustering techniques

Hierarchical clustering builds a cluster hierarchy called a dendrogram. Hierarchical clustering methods are categorized into agglomerative (bottom-up) and divisive (top-down) ones [13–18]. An agglomerative clustering starts with one-point cluster and recursively combines the most appropriate clusters. Divisive clustering considers the data set as one big cluster and gradually divides it into smaller ones. The main disadvantage of this clustering method is that the implementation of the criterion of finishing the clustering process is rather vague. The examples of this type of clustering are the following algorithms: SLINK, COBWEB, CURE, and CHAMELEON.

While hierarchical algorithms create clusters slowly, partitioning algorithms study clusters directly trying to identify them as areas highly populated with data. The latter are categorized into probabilistic clustering (EM framework, SNOB, AUTOCLASS, MCLUST), k-medoids methods (algorithms PAM, CLARA, CLARANS), and k-means methods [13–18]. The disadvantage of probabilistic clustering is that clear identification of the number of clusters is probabilistic. The disadvantages of two other methods are following: the initial choice centers (as medoids) exist in the data set (for k-medoids), and the result of clustering strongly depends on the initial guess of centroids (for k-means).

Density based partitioning algorithms try to determine densely connected components of data set. Density-based connectivity is used in the algorithms DBSCAN, OPTICS, DBCLASD [13–18].

In spite of the diversity of algorithms for data clustering, the density based approach was used in this article. For estimating the data density the bivariate kernel density estimation has been used. The kernel density estimation helps to evaluate the probability density function of a random data set.

5. Algorithm description

5. 1. Input data preprocessing

As it was mentioned above, the main purpose of research is determining the number of clusters, namely, we are interested in whether there is more than one cluster. For the explanation of the method operation the input data were selected to form a sample consisting of 90 points. The data were selected randomly as two-dimensional array [x, y] with uniform distribution. These points have been pre-selected for demonstrating three clusters. This is done so that the person could estimate the number of clusters and compare them with the number of clusters obtained by this algorithm. The initial data could be generated in the Matlab code, but in this case they were read off from the Excel document. The distribution of the initial data is shown in Fig. 1.

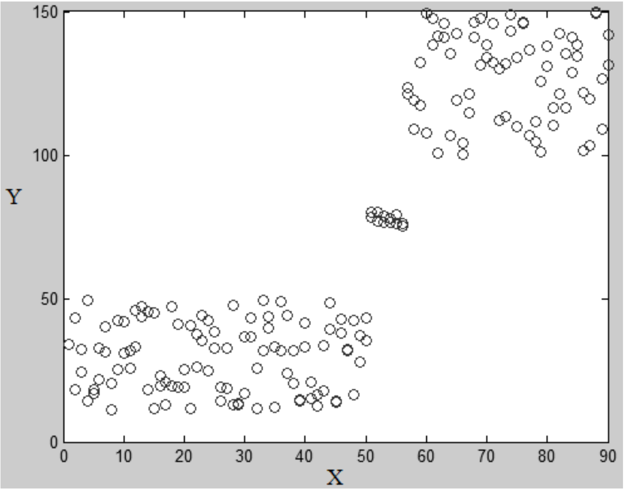


Fig. 1. Example of initial data

For the received pointes, the largest density of accumulations has been found using the built-in Matlab toolbox called Kernel density estimation toolbox (KDE). The KDE analysis is a general Matlab class for k-dimensional kernel density estimation. In our case of 3 dimensional data, this density function was modified. The initial density distribution of points is shown in Fig. 2, 3.

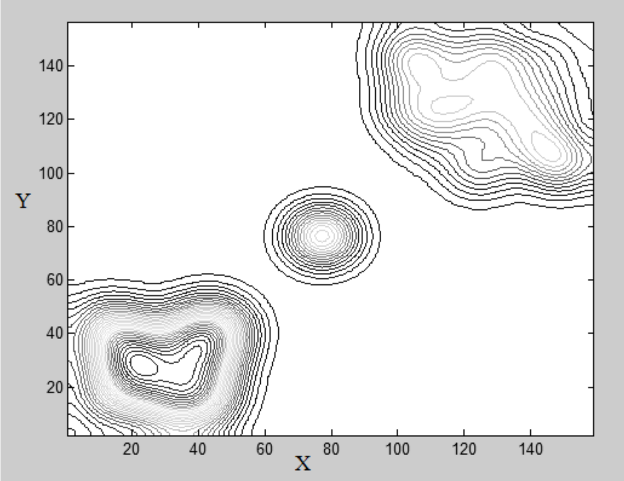


Fig. 2 Density distribution of initial data using 2-D figuration

Fig. 2 shows the regions whose area depends on the scatter of points, and the intensity shows the change in the density distribution of these points. The more the color is saturated the larger is the density of these points in this region.

Fig. 3 shows the density distribution of the points in 3-dimentional image. Density distribution of the points has been expressed clearly by local maximums (peaks).

The next step is finding all possible peaks. Each peak is a smooth point spread function (PSF). In reality, there is always noise, which typically has a one pixel variation. Because the peak spread function is assumed to be larger than one pixel, the true local maximum of that PSF can be obtained if we can get rid of these single pixel noise variations. Medfilt2, which is a 2D median filter, is used here for eliminating this noise. Next we smooth the image using conv2, so that with high probability there will be only one pixel in each peak that will correspond to the true local maximum

PSF. Medfilt2 and conv2 are Matlab functions. The result of peak finding is shown in Fig. 4.

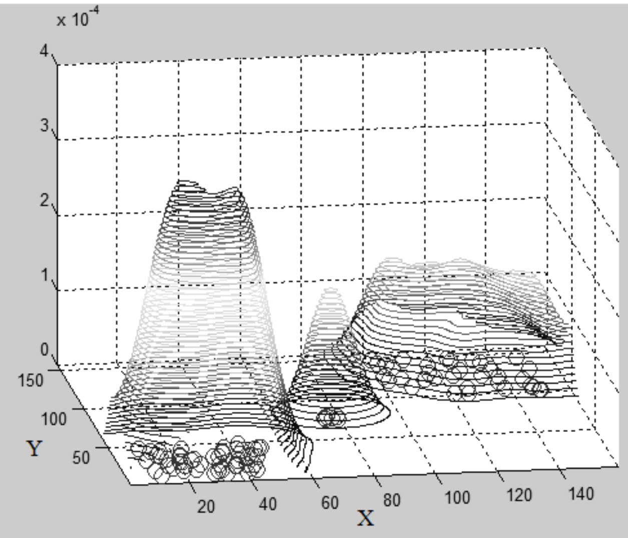


Fig. 3 Density distribution of initial data using 3-D figuration (3-D view of Fig. 2)

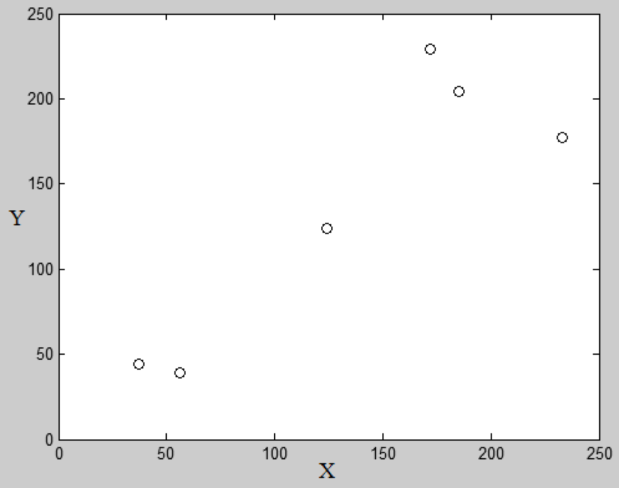


Fig. 4. Example of found local maximums (from Fig. 3)

The results in Fig. 4 show that peaks are located in three existing regions. One peak is located in the middle of the image, two peaks are in the bottom region, and three of them are located at the top of image. Let us denote the number of peaks as $p_{i,j}$, where i=1..I is the step by step change of the density distribution of the points, j=1..J is the designation of each region at each change of a step. Let us also denote the height of local maximums $h_{i,j}$, and in turn the height of k-th local maximum as $h^{k}_{i,j}$.

The total number of found peaks is denoted as $p_{1,j}$ at the first step, and it is equal to 6. The corresponding peak heights and their coordinates are shown below:

$h^{1}_{1,1}$ (37, 44)=0,0004;

$h^{2}_{1,1}$ (56, 39)=0,0003;

$h^{3}_{1,2}$ (124, 124)=0,0002;

$h^{4}_{1,3}$ (172, 229)=0,0001;

$h^{5}_{1,3}$ (185, 204)=0,0001;

$h^{6}_{1,3}$ (233, 177)=0,0001.

Next step is the replacement of these regions to a black-white image which is shown in Fig. 5.

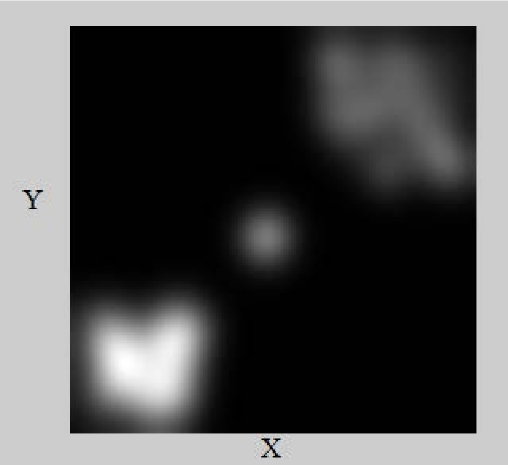


Fig. 5. Example of monochrome regions (from Fig. 2)

Filtering has been used to find the edges of the regions (Fig. 6). In our case the filtering is done via a canny filter. Edge detection is an image processing technique for finding the boundaries of objects within images. It works by detecting discontinuities in brightness. The canny method finds edges by looking for local maxima of the gradient of binary image. The method uses two thresholds, to detect strong and weak edges, and includes the weak edges in the output only if they are connected to strong edges. This method is therefore less likely than the others to be fooled by noise, and more likely to detect true weak edges. We have chosen the sensitivity thresholds for the canny method so that identifying the edges could be the best.

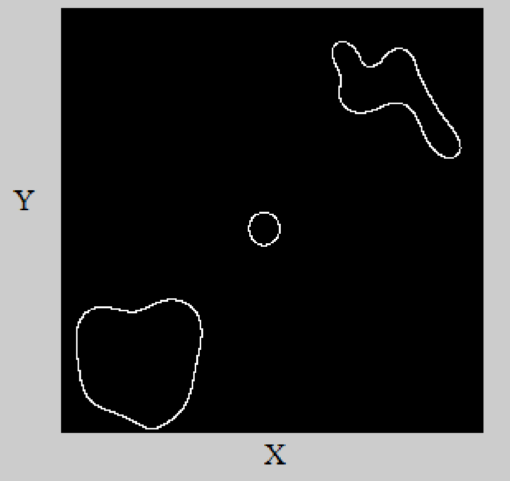


Fig. 6. Example of finding the edges

The choice of optimal sensitivity thresholds for the canny method depends on the density distribution of the pointes.

The next step is filling the allocated regions that are presented in Fig. 7.

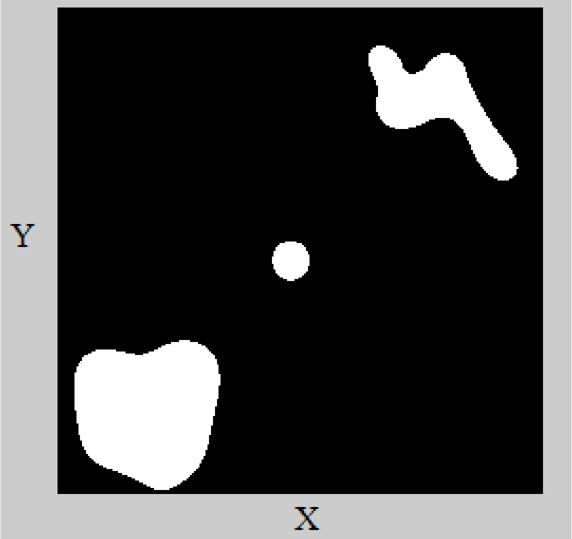


Fig. 7. Whitening the edges

The different shades of gray have been used to replace each white region. This is done in order to display the filled region correctly and to count the proper number of the points located in a given region. The selected regions and the points caught in them are shown in Fig. 8.

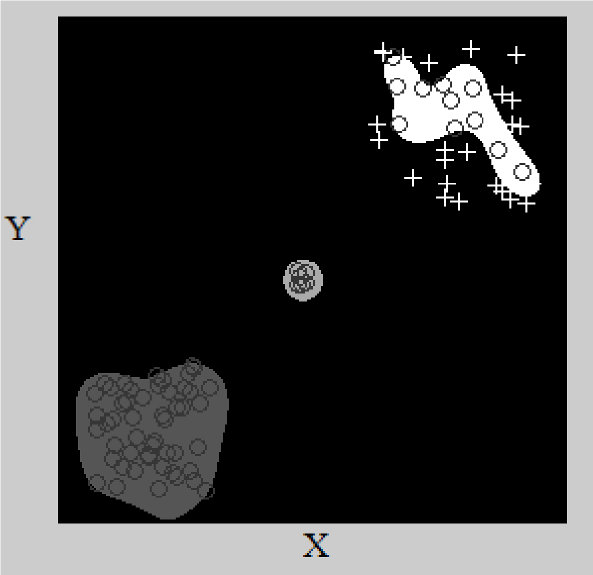


Fig. 8. Example of catching the points

Fig. 8 shows that not all points of a given density distribution are included to the selected region. Points belonging to a given region are marked with circles, and the points that are not included in it are marked with white crosses. So, in the presence of the initial density distribution of points and the region with high density created on the basis of this distribution, the area of these regions can be changed and the number of points in each region can be easily counted.

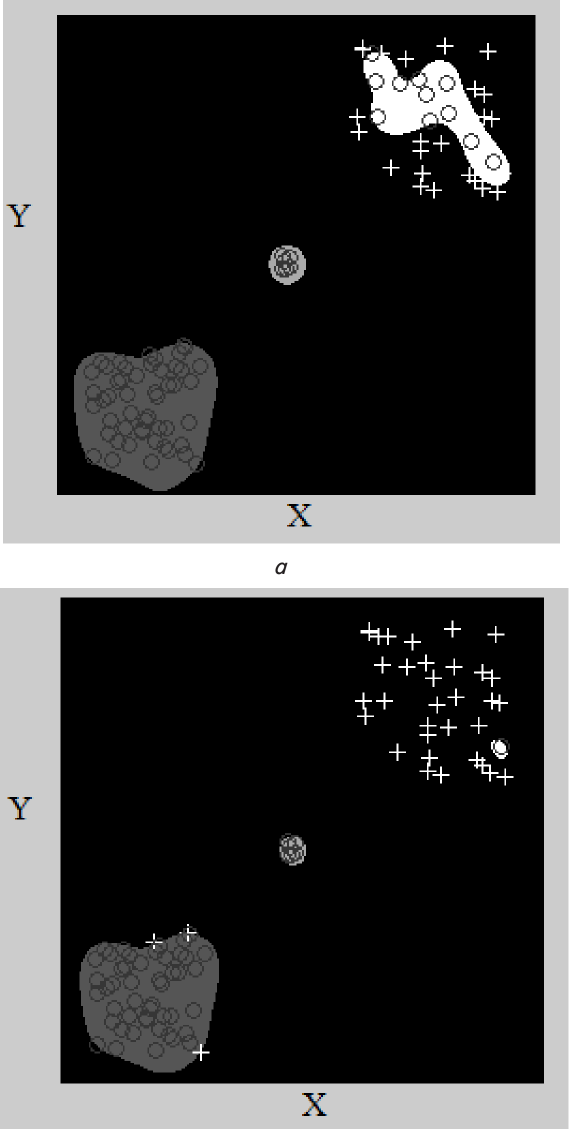
5. 2. Decision rule for detecting the number of clusters

The part of an algorithm which is shown above is a precondition for the creation of the decision rule for data clustering. This decision rule is based on the initial density distribution of the points and the number and magnitude of the peaks at the found regions. Making a decision about the number of found clusters follows the directions below:

– Specify the initial data for the approach, denote the initial number points as $n_{1,j}$

– Proceeding from this, find the number $p_{i,j}$ and height $h_{i,j}$ of all peaks in the found regions.

– At every change of the step, find the number of points that belong to each of the found regions (Fig. 9).



b

Fig. 9. Comparison of points caught in the given region, by changing the step: *a* — initial density; *b* — density decreasing

Knowing a priori the number of points in our data set by finding all peaks and its heights, as well as the percentage of points that belong to these regions, we can conclude whether this region is a cluster. The decision rule is presented below at Fig. 10.

$$\begin{cases} \text{if}\left(n_{i>1,j} > \dfrac{n_{1,j}}{2}\right) \quad \text{OR} \\ \left[\left(n_{i>1,j} \geq 0.05 n_{1,j} \quad \text{AND} \quad n_{i>1,j} < \dfrac{n_{1,j}}{2}\right) \text{AND} \left(p_{i,j} > 1 \quad \text{OR} \quad h^{k}_{i,j} > h^{\min}_{i,j}\right) \text{AND} \left(\dfrac{n_{i>1,j}}{n_{1,j}} \geq 0.09\right)\right] \text{then}\, Cl_j = 1 \\ \text{else} \quad Cl_j = 0 \end{cases}$$

Fig. 10. Decision rule for designed clustering algorithm

Where:

S is the initial data set;

S_i is the step by step change the value of S, where i=1..I is the step;

$S_{i,j}$ is the value of S_i, when changing the step i and specifying the sequence of found regions j, where j=1..J;

$n_{i,j}$ is a number of found pointes in the region $S_{i,j}$;

$p_{i,j}$ is a number of local maximums (peaks);

$h_{i,j}$ is the height of local maximum (peaks);

$h^{k}_{i,j}$ is the height of k-th local maximum, where k=1.. $p_{i,j}$;

$h^{\min}_{i,j}$ is the height of the smallest local maximum (peak);

Cl_j=1 and Cl_j=0 are existence and nonexistence of the cluster in the region j respectively.

The $n_{i,j}$ is a number of all points that belong to each found region. If the region contains less than 5 % of the total number of initial points $n_{1,j}$, it is not considered to be a cluster.

So knowing all the necessary parameters we can summarize. At the first step let us find the areas of all regions with the maximum density distribution (Fig. 6). In our case there are 3 such regions. In the region $S_{1,1}$, $n_{1,1}$=55,5 % of all pointes $n_{1,j}$ are located. In conclusion, this region is a cluster, that is, Cl_1=1.

The region $S_{1,2}$ is much less than the previous one (Fig. 6). In this region there are $n_{1,1}$=6,66 % of all pointes $n_{1,j}$. And also at low density distribution the height of the peak $h^{3}_{1,2}$ is significant.

The step being changed (i=1..I), the number of points situated in this region changes slightly. Basing on this fact, it is possible to say that this region is a cluster, Cl_2=1.

The third found area $S_{1,3}$ is the second-largest. Despite it, the percentage of points belonging to it is only $n_{1,3}$=12,2 %. There are also three peaks in it: $h^{3}_{1,2}$, $h^{3}_{1,2}$,$h^{3}_{1,2}$, the heights of which are lower than in the previous regions.

Changing the step i and reducing the density distribution, in some time, causes the disappearance of the third region, though the other two regions still exist (Fig. 7). The third region is not a cluster, because the condition of changing the regions is not executed and this means that Cl_3=0.

This decision rule is one of the major parts of the designed clustering algorithm. Presented clustering algorithm as well as decision rule is shown at a diagram which is below at Fig. 11.

This diagram illustrates the way of processing the input data up to the stage of identifying the number of clusters.

6. Validation of presented algorithm

For the simulation of this algorithm operation other data were taken. Data were generated randomly with uniform distribution. The size of the initial data set is the same for two cases, that is, the number of points is equal to 90. Based on the a priori density distribution of the points one large region is built (Fig. 12, *a*), and in Fig. 12, *b* there are two regions.

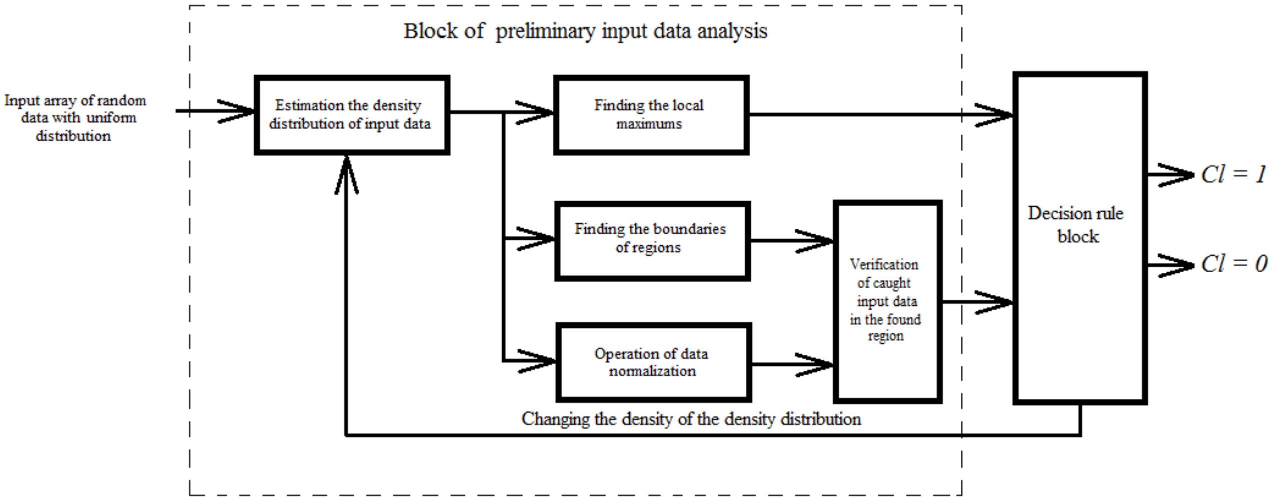
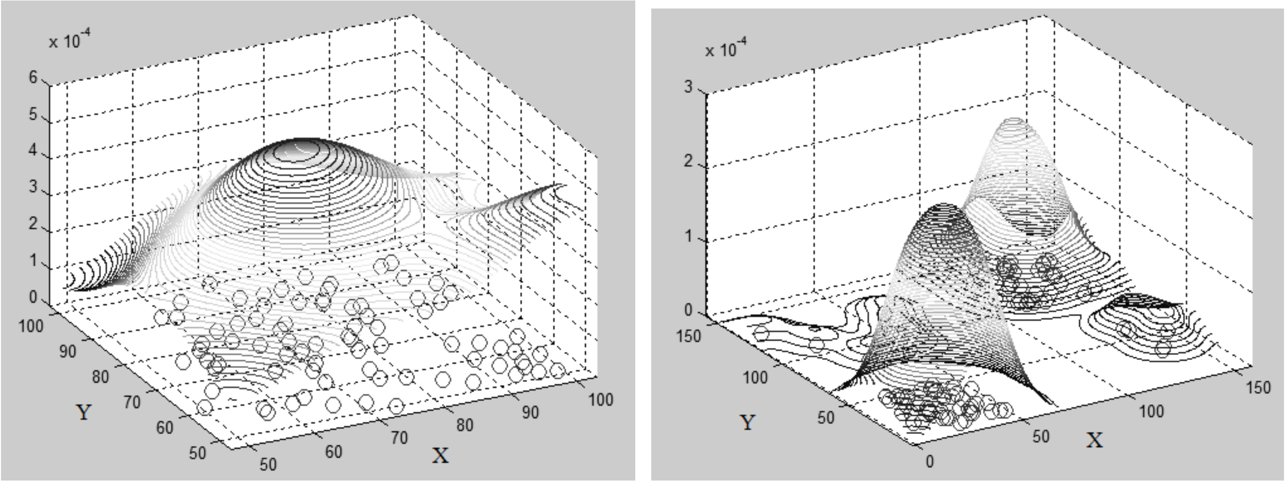

Block of preliminary input data analysis
Input array of random data with uniform distribution
Estimation the density distribution of input data
Finding the local maximums
Finding the boundaries of regions
Operation of data normalization
Verification of caught input data in the found region
Decision rule block
Cl = 1
Cl = 0
Changing the density of the density distribution

Fig. 11. The diagram of designed clustering algorithm



a *b*

Fig.12. Initial data: *a* — case 1; *b* — case 2

The result of the simulation shows that one cluster has been found in Fig. 12, *a*, and in the other case two clusters have been found in Fig. 12, *b*. These results confirm the correctness of the proposed clustering algorithm.

7. Conclusion

As a result of research, the data set clustering algorithm has been developed. The data set is the only initial parameter which is required, so this is the advantage of this algorithm, at the same time all paper set up research objectives have been fulfilled. Developed algorithm can be compared with the existing density based clustering algorithms and presented in tabular form at Table 1.

Table 1

Comparison of density-based clustering algorithms

Algorithm	Input parameter	Arbitrary shape	Varied density
DBSCAN	Radius and Minpts	+	–
FDBSCAN	Radius and Minpts	+	–
ODBSCAN	Number of identical circles, Radius and Minpts	+	–
VDBSCAN	Number of clusters	+	+
ST–DBSCAN	Three parameters are given by user	+	–
Inc. DBSCAN	Radius and Minpts	+	+
RDBC	Defining values	+	–
Developed algorithm	*Input data set*	+	+

To make a decision in this algorithm three parameters have been used, namely: the number of points belonging to a particular region at a particular step, the number of peaks and their height. Work on improving the algorithm will be continued in further research.

References

1. Kudo, M. Comparison of algorithms that select features for pattern classifiers [Text] / M. Kudo, J. Sklansky // Pattern Recognition. – 2000. – Vol. 33, Issue 1. – P. 25–41. doi: 10.1016/S0031-3203(99)00041-2
2. Wernick, M. N. Machine Learning in Medical Imaging [Text] / M. N. Wernick, Y. Yang, J. G. Brankov, G. Yourganov, S. C. Strother // IEEE Signal Processing Magazine. – 2010. – Vol. 27, Issue 4. – P. 25–38. doi: 10.1109/msp.2010.936730
3. Solomon, C. J. Fundamentals of Digital Image Processing: A Practical Approach with Examples in Matlab [Text] / C. J. Solomon, T. P. Breckon. – Wiley-Blackwell, 2010. – 328 p. doi: 10.1002/9780470689776
4. McCallum, A. Efficient Clustering of High Dimensional Data Sets with Application to Reference Matching [Text] / A. McCallum, K. Nigam, L. H. Ungar // Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining, 2000. – P. 169–178. doi: 10.1145/347090.347123
5. Deepti, S. Clustering Techniques: A Brief Survey of Different Clustering Algorithms [Text] / S. Deepti, S. Lokesh, S. Sheetal, S. Khushboo // International Journal of Latest Trends in Engineering and Technology (IJLTET). – 2012. – Vol. 1. – P. 82–87.
6. Khushali, M. NDCMD: A Novel Approach Towards Density Based Clustering Using Multidimensional Spatial Data [Text] / M. Khushali, A. Swapnil, M. Sahista // International Journal of Engineering Research & Technology (IJERT). – 2013. – Vol. 2, Issue 6.
7. Shou, S.-G. A Fast DBSCAN Algorithm [Text] / S.-G. Shou, A.-Y. Zhou, W. Jin, Y. Fan W.-N. Qian // Journal of Software. – 2000. – P. 735–744.

8. Peter, J. H. An Optimised Density Based Clustering Algorithm [Text] / J. H. Peter, A. Antonysamy // International Journal of Computer Applications. – 2010. – Vol. 6, Issue 9. – P. 20–25. doi: 10.5120/1102-1445
9. Wei, W. Improved VDB scan with global optimum K [Text] / W. Wei, Z. Shuang, R. Bingfei, H. Suoju. – 2013.
10. Birant, D. ST-DBSCAN: An algorithm for clustering spatial-temporal data Data Knowl [Text] / D. Birant, A. Kut // Data & Knowledge Engineering. – 2007. – Vol. 60, Issue 1. – P. 208–221. doi: 10.1016/j.datak.2006.01.013
11. Navneet, G. An Efficient Density Based Incremental Clustering Algorithm in Data Warehousing Environment [Text] / G. Navneet, G. Poonam, K. Venkatramaiah, P. C. Deepak, P. S. Sanoop // 2009 International Conference on Computer Engineering and Applications IPCSIT. – 2011. – Vol. 2.
12. Rehman, M. Comparison of density-based clustering algorithms [Electronic resource] / M. Rehman, S. A. Mehdi. – Available at: https://www.google.com.ua/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&ved=0CBwQFjAA&url=http%3A%2F%2Fwww.researchgate.net%2Fprofile%2FSyed_Atif_Mehdi%2Fpublication%2F242219043_COMPARISON_OF_DENSITY-BASED_CLUSTERING_ALGORITHMS%2Flinks%2F5422e1120cf26120b7a6b36e.pdf&ei=LHgRVaSTA6Gv7Abh34CACw&usg=AFQjCNFA9JnzuIbam4BOKYCS_30Yw8Czmg&sig2=wNiTYQiNzFKcDOfEV3mLFw&cad=rja
13. Berkhin, P. Survey Of Clustering Data Mining Techniques [Electronic resource] / P. Berkhin. – 2002. – Available at: http://www.cc.gatech.edu/~isbell/reading/papers/berkhin02survey.pdf
14. Abu Abbas, O. Comparison Between Data Clustering Algorithms [Text] / O. Abu Abbas // The International Arab Journal of Information Technology. – 2008. – Vol. 5, Issue 3. – P. 320–325.
15. Gan, G. Data Clustering: Theory, Algorithms, and Applications [Text] / G. Gan, M. Chaoqun, W. Jianhong. – ASA-SIAM Series on Statistics and Applied Probability, SIAM, Philadelphia, ASA, Alexandria, 2007. – 466 p. doi: 10.1137/1.9780898718348
16. Jiawei, H. Data Mining: Concepts and Techniques. Second Edition [Text] / H. Jiawei, M. Kamber, J. Pei. – Series Editor Morgan Kaufmann Publishers, 2006. – 800 p.
17. Riley, K. F. Mathematical methods for physics and engineering [Text] / K. F. Riley, M. P. Hobson, S. J. Bence. – Cambridge University Press, 2010. – 1359 p.
18. Anil, K. J. Algorithms for clustering data [Text] / K. J. Anil, R. C. Dubes. – Prentice-Hall, Inc. Upper Saddle River, NJ, USA, 1988.

УДК 681.32

DOI: 10.15587/1729-4061.2015.39947

ОЦЕНКА КОРРЕКТИРУЮЩЕЙ СПОСОБНОСТИ ЦИКЛИЧЕСКИХ КОДОВ НА ОСНОВЕ ИХ АВТОМАТНЫХ МОДЕЛЕЙ

В. П. Семеренко
Кандидат технических наук, доцент
Кафедра вычислительной техники
Винницкий национальный
технический университет
Хмельницкое шоссе, 95,
г. Винница, Украина, 21021
E-mail: VPSemerenko@ukr.net

Проведено дослідження коректувальної здатності різних підкласів циклічних кодів з використанням кінцевих автоматів в двійкових полях Галуа – лінійних послідовнісних схем (ЛПС). Показано, що структура нульових циклів ЛПС однозначно визначає кількість випадкових помилок і пакетів помилок, які виявляються та виправляються. Введені нові характеристики коректувальної здатності циклічних кодів

Ключові слова: циклічні коди, кодова відстань, коректувальна здатність коду, лінійна послідовнісна схема

Проведено исследование корректирующей способности различных подклассов циклических кодов с использованием конечных автоматов в двоичных полях Галуа – линейных последовательностных схем (ЛПС). Показано, что структура нулевых циклов ЛПС однозначно определяет количество обнаруживаемых и исправляемых случайных ошибок и пакетов ошибок. Введены новые характеристики корректирующей способности циклических кодов

Ключевые слова: циклические коды, кодовое расстояние, корректирующая способность кода, линейная последовательностная схема

1. Введение

Развитие средств связи и растущий спрос на телекоммуникационные услуги требует дальнейшего улучшения качества таких услуг для пользователей: увеличения пропускной способности (числа абонентов в случае мобильной связи), повышения достоверности передачи, снижения потребляемой мощности аппаратурой.