A Two Stage Clustering Method Combining Self-Organizing Maps and Ant K-Means

Jefferson R. Souza, Teresa B. Ludermir, and Leandro M. Almeida

Center of Informatics, Federal University of Pernambuco
Av. Prof. Luis Freire, s/n, Cidade Universitária Recife/PE, 50732-970, Brazil
{jrs2,tbl,lma3}@cin.ufpe.br

Abstract. This paper proposes a clustering method SOMAK, which is composed by Self-Organizing Maps (SOM) followed by the Ant K-means (AK) algorithm. SOM is an Artificial Neural Network (ANN), which has one of its characteristics, the nonlinear projection from a high dimensionality of the sensorial space. AK is based in the Ant Colony Optimization (ACO), which is a recently proposed meta-heuristic approach for solving hard combinatorial optimization problems. The AK algorithm modifies the K-means on locating the objects and these are then clustered according to the probabilities which in turn are updated by the pheromone. The SOMAK has a good performance when compared with some clustering techniques and reduces the computational time.

Keywords: Self-Organizing Maps, Ant Colony Optimization and Unsupervised Learning.

1 Introduction

With the substantial reduction of data storage cost, a great improvement in the performance of computers and the popularization of computer nets, a great amount of data information is being produced every day everywhere. So, a great quantity scale of databases has created the necessity of developing some techniques of data processing useful for the clustering of data or data mining [4].

The K-means algorithm is the most commonly used partitive clustering algorithm because it can be easily implemented and it is very efficient in terms of the execution time. The major problem with K-means it is the definition of k to the clustering problem. SOM [6] is an ANN which allows the visualization of high dimensionality data and also implements an ordered mapping of a distribution of high dimension within a regular grid of low dimension. This ordered grid can be used as a convenient visualization to show different characteristics of the SOM. The algorithm SOMK, that is SOM followed by K-means, does not need to define the ideal number of k-clusters [12]. The algorithm AK is based on the Ant Colony Optimization (ACO) [3], recently proposed for solving hard combinatorial optimization problems. We used AK to find the optimized k [7]. The advantage of the proposed algorithm SOMAK is that it needs less time with small number of clusters to solve a problem. We have to point out that the aim of

C. Alippi et al. (Eds.): ICANN 2009, Part I, LNCS 5768, pp. 485–494, 2009.

this paper is not to find an optimal clustering for the data, but to obtain a view about the structure of data clusters using SOMAK, besides trying to reduce the number of clusters and the computational time.

In the case of the experiments, a comparison between the results of direct data (SOM and K-means) and the clusters of prototypes vectors of SOMK and SOMAK is carried out. This paper contains eight sections. Section 1 describes our research motivation. Section 2 contains the related work. Section 3 contains the methods of clustering. In section 4, a two-stage method combining SOM and Ant K-means is described. Section 5 shows the material and methods. Section 6 describes the experimental results and discussion. Section 7 contains the conclusion and future work. Finally, the acknowledgement are presented.

2 Related Works

Kuo et al. used AK in the analysis of clusters [7]. The algorithm AK modifies the k-means locating the objects and these are then clustered according to the probabilities which in turn are updated by the pheromone according to the total within cluster variance (TWCV). The experimental results showed that AK is better than the other two methods, SOMK and SOM followed by the genetic k-means algorithm [7]. The only problem for AK is that the number of clusters is required, that is, it is necessary to give the number of clusters to algorithm AK for it to be started.

Vesanto and Alhoniemi combined SOM and K-means [12] to solve the clustering problem. Particularly, the use of hierarchical agglomerative clustering and the partitive clustering using K-means are investigated. The procedure consists of two stages, firstly using a SOM to produce the prototypes, which are then clustered in the second stage by the K-means. The results of the clustering using a SOM as an intermediary phase was computationally effective, besides comparing the results directly obtained from the data, considering the original difficulties from the properties of the K-means algorithm. Trying to solve the needs of the algorithms which were seen and described above, we need to develop some useful techniques of data processing to improve the solution of the data clustering or data mining. So, this paper proposes a method of clustering based on two stages combining SOM and Ant K-means for the analysis of clusters.

3 Methods of Clustering

3.1 K-Means

The K-means method of clustering is one of the simplest algorithms of unsupervised learning to solve the clustering problem. The aim is to divide the data set within k clusters fixed a priori. The algorithm consists of two stages: an initial stage and an iterative stage. The initial stage involves the definition of the k centroids, one for each cluster. The second iterative stage repeats the signature of each point of data for the closest centroid and k new centroids are calculated according to the new signature [10]. This interaction stops when a certain criterion is found; for example, number of interactions. Given a set *nPat*, suppose we want to classify the data within k groups, the algorithm tends to minimize a function of error, such as a mean squared error defined as: $E = \sum_{k=1}^{C} \sum_{i=1}^{nPat} ||x_\imath - c_k||^2$. Where C represents the number of clusters, *nPat* the number of samples, x the entry of each sample and c_k is the center of cluster k.

3.2 Ant Colony Optimization

The ACO was proposed by Dorigo [3]. When we refer to the colonies of ants, we observe the ants communicate to each other just in an indirect form in their environment by the substance called pheromone. Paths with higher levels of pheromone will be likely to be chosen and consequently reinforced while the intensity of pheromone along the paths that are not chosen is reduced by evaporation. Additionally, evaporation causes the pheromone level of all trails to diminish gradually. Hence, trails that are not reinforced gradually lose pheromone and will in turn have a lower probability of being chosen by subsequent ants. Evaporation is accomplished by diminishing the pheromone level of each trail by a factor ρ. Typical values for this evaporation factor ρ lie in the range [0.8, 0.99][3]. This is an important mechanism to update the pheromone on the trails according to $\tau_{ij} \leftarrow (1 - \rho) * \tau_{ij} + \frac{Q}{TWCV}$, whose parameters are explained in more details in the next section. This form of indirect communication is known and gives the colony of ants the capacity for finding the shortest path [8]. There are some works related to the algorithms of clusters based on ACO. Yuqing et al. Proposes algorithm of K-means clusters based on density and on the Colony of Ants [13]. This algorithm is a new K-means algorithm based on the density and theory of ants, which solved the problem of the local minimum by the random ants, besides manipulating the initial parameters of K-means. Handl et al. proposes clustering based on ants [5].

3.3 Ant K-Means

The choice of this algorithm is because it produced satisfactory results regarding the clustering problem. In this method, It is necessary to provide the number of clusterings like in the conventional K-means algorithm for AK algorithm. Suppose $E = O_1, O_2, ..., O_n$ the set of n data or objects, where O represents the objects collected from the database, in that each object has k attributes, where $k > 0$. Bellow some important parameters such as: α: The relative importance of the trail: $\alpha \geq 0$; β: The relative importance of the visibility: $\beta \geq 0$; ρ: The pheromone decay parameter: $0 < \rho < 1$; Q: A constant; n: Number of objects; m: Number of ants; nc: Number of clusters; T: is the set includes used objects. The maximal number recorded by T array will be n, i.e., $T = O_a, O_b, ..., O_t$ where *a*, *b*, ..., *t* are the points that ant has been. T_k: The set T is performed by ant k. $O_{center}(T)$: The object which is the center of all objects in T, i.e., $O_{center}(T) = \frac{1}{nT} \sum_{i=1}^{n} O_i$, where nT is the number of objects in T. $TWCV$: total within cluster variance, i.e., $\sum_{k=1}^{nc} \sum_{i=1}^{n} (O_i - O_{center(T_k)})^2$.

The Algorithm 1 shows the procedure Ant K-means in details above.

Algorithm 1. The procedure of Ant K-means [7]

```
Procedure Perturbation: Each Ant starts at random object and chooses the centroid randomly of cluster to move for all Ant k. Calculating $O_{center(T_k)}$ where $k = 1, 2, ..., nc$ and TWCV.
Procedure Ant K-means: Input the number of clusters and the corresponding centroids, and set the parameter $\alpha$, $\beta$, $\rho$, number of iterations and ants. Lay equal pheromone on each path.
while the number of iterations is not reached do
    while TWCV is not changed do
        Updating pheromone by $\tau_{ij} \leftarrow (1 - \rho) * \tau_{ij} + \frac{Q}{TWCV}$.
        Each Ant k chooses the centroid to move with P, i.e.,
        $P = \frac{\tau_{kc}^{\alpha} * \eta_{kc}^{\beta}}{\sum_{i=1}^{nc} (\tau_{ki}^{\alpha} * \eta_{ki}^{\beta})}$
        Calculate $O_{center(T_k)}$ where $k = 1, 2, ..., nc$
        Calculate TWCV (Total Within Cluster Variance).
    if TWCV is smaller than the smallest TWCV then
        replace it.
    else
        Pertubation
```

3.4 SOM-Based Two-Stage Methods

A proposed method of clustering based on two stages is useful to improve the main disadvantages of a partitive method of clustering; for example, K-means due to its sensitivity to the initial prototypes and the difficulty in determining a proper number of k clusters. Generally, a SOM-based two-stage method has two possible forms of working. In the first one SOM is initially used to determine the number of groups and the center of the initial groups for the Ant K-means. The initial center of a group can be obtained from the weight vector corresponding to the center of the groups on the topology of SOM net. In the second form, the initial maps of SOM net present a large set of scale data on its topology and generates the topological coordinates of the prototypes for future clusters in the second stage. The method used in the second phase is the Ant K-means procedure. The main advantage of a SOM-based two-stage method is the reduction of the computational time by the hierarchical clustering method or partitive for the large and complex sets of data [12].

To show this characteristic, it was necessary to train a SOM net by using the algorithm of sequential training for the data set [1]. The maps were trained in two phases: a rough training with width of initial neighborhood $\sigma_1(0)$ and big learning rate and another phase called fine-tuning with width of initial neighborhood

[1] Training of SOM net freely available in the package Matlab SOM Toolbox which was used in the implementation of the proposed method. For further information, see URL http://www.cis.hut.fi/projects/somtoolbox/

$\sigma_2(0)$ and small learning rate, which the width of the neighborhood decreases linearly to 1.

4 SOMAK

The method proposed in this paper, SOMAK, can be seen in Fig. 1. SOMAK uses SOM net as a classifier of characteristics about the entry data instead of clustering the data directly. First, a large set of prototypes is formed by using SOM. The prototypes can be interpreted as “proto-clusters”, which are in the next step combined to form the true clusters. Each data vector of the original data set belongs to the same cluster like its closest prototype. In the present study, the number of clusters and the centroid of each cluster are generated from SOM net. In order to validate the solution of clustering analysis, the framework Monte Carlo [9] was used in this paper. SOMAK uses SOM to determine the initial points and then uses the Ant K-means procedure to find out the final solution, i.e., AK to determine the number of clusters. The benefit of this approach is the reduction of the computational cost. The second advantage is the reduction of the clusters size. The reduction of the noise is another benefit. The prototypes are the local mean of the data and so, less sensitive to the random variations than the original data.

For this reason, it is convenient to cluster a set of prototypes, instead of the data directly [12]. Consider N samples of the data using Ant K-means algorithm which is described in section 3.3. This involves to make attempts of clustering with different values for the number of prototypes which were obtained by SOM net. The computational time is proportional to the $\sum_{k=2}^{C_{max}} Nk$, where C_{max} is the pre-established maximum number of clusters and k represents the number of initial clusters. When a set of prototypes is used in an intermediary step (Fig. 1 - 1st level of abstraction), the total time is proportional to $NM + \sum_{k=1}^{Mk}$, where M is the number of prototypes obtained. With $C_{max} = \sqrt{N}$ and $M = 5\sqrt{N}$, the reduction of the computational time is based on $\frac{\sqrt{N}}{15}$ or about six-fold for

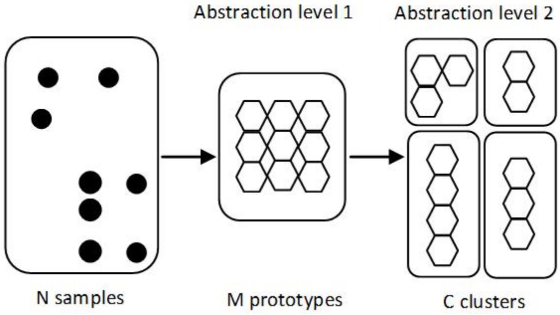

Fig. 1. The first level of abstraction is obtained through the creation of a set of prototype vectors by using SOM. Algorithm SOMAK creates the second level of abstraction carrying out the cluster of M prototypes [12].

$N = 10000$ [12]. In our case, we used ten-folds and $N = 1000$ for the carrying out of the experiments. Evidently this is a very rough estimate, since it is an estimate over the other; and many practical and experimental considerations are ignored.

5 Material and Methods

The carried out experiments were: synthetic data, real data, the method Monte Carlo, to check the efficiency of the four clustering methods. To carry out the experiments it was used a machine Intel(R) Core (TM) 2 Quad, processor 2.40GHz, memory RAM 3.00GB, operational system Microsoft Windows XP Professional version 2002 Service Pack 3.

5.1 Data Sets

In this paper, five data sets were used: *Lines*, *Banana*, *Highleyman* being these classified as synthetic data and *Contraceptive Method Choice* and *Glass* as real data.

The ***Lines*** basis consists of 1000 data points clustered in 10 segments. The other two bases of synthetic data [2] are arranged in the following way: A represents a data set of two classes in two dimensions; and N represents the number of samples of the vector generated with the number of samples by class. $N = [500, 500]$ having a total of 1000 data points. The ***Banana*** data basis shows too that the data points are distributed in a normal distribution in the form of a banana with standard deviation *S=1* in all directions. Now the ***Highleyman*** third set of data, besides the arrangements mentioned before, is divided into two classes: the ***1st class*** contains 500 data points for each one of the Gaussians with mean 1 and 0 and variances 0 and 0.25, the ***2nd class*** contains 500 data points for each one of the two Gaussians with mean of 0.01 and 0 and variances 0 and 4. The real data used the repository UCI [1]. The ***Contraceptive Method Choice or CMC*** represents the problem of predicting the choice of a woman's current contraceptive method based on her economic and socio-demographic characteristics. The number of instances is 1473, divided into three classes. The number of attributes is 10, including the class. The second real basis is the ***Glass***, this database has as an objective to determine if some glass belongs to a kind "float" or not. The study of classification of this kind of glass was motivated by a criminological investigation in which several tests were made about the glass. The number of instances is 214, divided into six classes. The number of attributes is 11, including the class. All the synthetic data and the partitions of the real data were obtained through random numbers (Monte Carlo). After that, it was used the stratified cross validations with ten-folds on the databases providing the training and test sets for all the clustering methods.

[2] Data sets, freely available in the package Matlab PRTools: Toolbox for Pattern Recognition was used. For further information, see URL
http://prtools.org/academic.html

So, it is reasonable to accept the reliability of the generator of random numbers. Finally, 30 executions about the project were carried out.

5.2 Parameters Setup

The parameters considered in this paper are those which affect direct or indirectly the clustering techniques, which were already described in the previously sections to solve the clustering problem. According to [3], there are several combinations to determine the parameters as applied to ant colony system. Normally, the parameters are $\alpha = 0, 0.5, 1, 2, 5$, $\beta = 0, 1, 2, 5$, $\rho = 0.3, 0.5, 0.7, 0.99, 0.999$ and $Q = 1, 100, 10000$. There are 300 combinations of parameters; the results showed in [2] that $\alpha = 0.5$, $\beta = 1$, $\rho = 0.9$ and $Q = 1$ in this method has the smallest variance, where m = 2 obtained the best results compared with m = 4 suggested by Marco Dorigo [3]. Table 1 shows the parameters of the clustering techniques.

Table 1. Main Parameters of Clustering Techniques

Clustering Techniques	Parameters
SOM	Attributes number = pattern quantity input, Lines size grid = 19, Columns = 17, Initial radius = 10, Final = 2, $\sigma_1(0) = 10$, $\sigma_2(0) = 2$, Initial learning rates were 0.5 and 0.05 respectively, Final = 0.99, Neighborhood function = Gaussian, Neighborhood format = hexa, Train type = epochs, Training size for rough phase = 3, fine-tuning phase = 10.
K-means	k = Number (n$^{\underline{o}}$) initial clusters, Initialize centers = k.
SOMK	k = SOM prototypes n$^{\underline{o}}$,Initialize centers = SOM centroids n$^{\underline{o}}$.
SOMAK	$\alpha = 0.5, \beta = 1, \rho = 0.9, Q = 1, n = 500,$ $m = 2, nc = SOMprototypesnumber.$

6 Experimental Results and Discussion

Then, to find the number of "proto-clusters" which obtains as a result 110 through SOM net; AK is used to cluster 500 data samples under the test set. Table 2 shows a comparison between SOMAK and SOMK to obtain a smaller number of clusters. The number of clusters and its centroids are obtained by SOM net and then uses AK to find the definite solutions. SOMAK has the best efficiency in comparison with SOMK, which is also the method composed of the two stages.

It is important to mention in Table 2 that the fact of the SOMAK method increases the number of clusters (compared to SOMK) does not mean to say that is bad, perhaps this increase may be necessary to have an improvement of entropy. SOMAK is applied as a technique of clustering for the case study because it obtained a smaller value of clusters. It will be presented the measure of Entropy, which showed a smaller value for SOMAK when compared to SOMK, the parameters Min, Max, Mea and Std represent respectively Minimum, Maximum, Mean and Standard Deviation in Table 3. The degree to which each cluster

Table 2. Results of the size of clusters obtained by the test set

Data sets	Initial Cluster	SOMK	SOMAK
Lines(I)	10	6	3
Banana(II)	2	7	4
Highleyman(III)	2	3	4
CMC(IV)	3	9	4
Glass(V)	6	5	3

Table 3. Results of the methods with 30 executions each to obtain the Entropy and the Computational Time (seconds)

Data sets	Methods	Results Entropy		Computational Time	
		Min—Max—Mea	Std	Min—Max—Mea	Std
I	SOM	0.043—0.154—0.103	0.028	1.764—1.811—1.791	0.012
	Kmeans	**0.003**—**0.026**—**0.006**	**0.005**	0.332—0.340—0.336	0.002
	SOMK	0.325—0.398—0.366	0.018	2.346—2.397—2.376	0.014
	SOMAK	0.229—0.341—0.273	0.026	1.931—1.999—1.970	0.016
II	SOM	**0.036**—**0.122**—**0.074**	**0.021**	1.764—1.847—1.818	0.014
	Kmeans	0.426—0.477—0.460	0.012	0.263—0.291—0.269	0.005
	SOMK	0.375—0.457—0.415	0.019	2.396—2.488—2.456	0.016
	SOMAK	0.267—0.421—0.340	0.041	1.939—2.035—1.995	0.019
III	SOM	**0.212**—**0.326**—**0.268**	0.029	1.795—1.837—1.819	0.010
	Kmeans	0.453—0.514—0.482	**0.013**	0.269—0.302—0.277	0.005
	SOMK	0.410—0.491—0.450	0.018	2.436—2.485—2.463	0.014
	SOMAK	0.250—0.448—0.365	0.057	1.955—2.019—1.990	0.015
IV	SOM	0.478—0.509—0.493	**0.007**	4.494—4.552—4.531	0.016
	Kmeans	0.397—0.450—0.413	0.012	0.292—0.308—0.300	0.005
	SOMK	0.407—0.467—0.442	0.013	5.267—5.333—5.304	0.019
	SOMAK	**0.175**—**0.314**—**0.245**	0.035	4.748—4.827—4.793	0.017
V	SOM	0.306—0.397—0.365	**0.020**	3.427—3.577—3.554	0.030
	Kmeans	0.286—**0.368**—0.331	0.023	0.333—0.351—0.341	0.004
	SOMK	0.295—0.370—0.337	**0.020**	3.945—4.107—4.080	0.032
	SOMAK	**0.238**—0.372—**0.317**	0.036	3.478—3.630—3.602	0.030

consists of objects of a single class. For each cluster, the class distribution of the data is calculated first, i.e., for cluster j we compute p_{ij}, the probability that a member of cluster i belongs to class j as $p_{ij} = \frac{m_{ij}}{m_i}$, where m_i is the number of objects in cluster i and m_{ij} is the number of objects of class j in cluster i. Using this class distribution, the entropy of each cluster i is calculated using the standard formula [14], $e_i = -\sum_{j=1}^{L} p_{ij} log_2 p_{ij}$, where L is the number of classes. The total entropy for a set of clusters is calculated as the sum of the entropies of each cluster weighted by the size of each cluster, i.e., $e = \sum_{i=1}^{K} \frac{m_i}{m} e_i$, where K is the number of clusters and m is the total number of data points.

The Std parameter reported in Table 3 presented a smaller value for most of the methods of clusterings except for SOMAK. So, Table 3 showed also a great variability pointed out in the standard deviation parameter, resulting in a disadvantage for the proposed SOMAK method. In the majority of the experiments, the SOMAK method showed a smaller entropy in the parameters of Min, Max and Mea, when compared with the methods of clustering SOM, K-means and SOMK for IV and V data sets. For the I, II and III data sets the SOMAK method presented a smaller entropy when compared with SOMK. Table 3 also shows computation time for all the techniques of clusterings used in the experiments. K-means has always been the quickest one computationally, because it is a simple algorithm or of only one stage. However, this same algorithm presented a high entropy seen in Table 3, when compared with the SOM, SOMK and SOMAK methods. SOMAK had a longer time than K-means and obtained more satisfactory results when compared with SOMK as well in reference to the entropy as to the computational time.

It was concluded that the experimental results are statistically independent according to the application of Test t (hypothesis test). It was applied as well for the entropy as for the computational time respectively seen in Table 3 and with 5% of significance degree it showed that SOMAK is better than SOMK.

7 Conclusion and Future Work

The aim of this paper was to propose a method of clustering, SOMAK, composed of two stages by combining SOM and Ant k-means. The SOMAK method is capable of reducing the size of clusters, by finding a good performance when compared with other techniques of clustering (SOM, K-means and SOMK) and also capable of reducing the computational time of the experiments.

The algorithms of clusters described before were tested as well for the data directly as for the data trained by SOM net. It was used a SOM net as an intermediary step besides carrying out a comparison of the results obtained directly from the data. The results for the data generated by the Monte Carlo method showed that SOMAK is better than SOMK, because there was a reduction of the size of the clusters for the test set (Table 2), for it to have formed a better performance when compared with SOMK seen (Table 3) and finally, Table 3 also shows that SOMAK reduced the computational time in comparison with SOMK to solve the problem of data clusterings. So, the proposed method is a robust method of clustering. It can be applied to a lot of different kinds of clustering problems or combined with some other techniques of data mining to obtain more promising results.

For future works, the idea is readjusting the SOMAK algorithm with the purpose of reducing its computational time when compared with the methods described in this paper. The first method that will be observed is the ABSOM [11]. This one has better performance than SOM and also works very well in the analysis of clustering in two stages when it is used as a technique of preprocessing. So, this method is composed of two stages for the analysis of data, where it has demonstrated to be useful and effective.

Acknowledgments

This work is supported by CNPq and FACEPE (Brazilian research agencies).

References

1. Aha, D.: UCI machine learning repository, `http://archive.ics.uci.edu/ml/` (access January 16, 2009)
2. Berkhin, P.: Survey of Clustering Data Mining Techniques, Accrue Software, `http://www.accrue.com/` (access January 7, 2009)
3. Dorigo, M., Stützle, T.: The Ant Colony Optimization Metaheuristic: Algorithms, Applications, and Advances, Technical Report IRIDIA (2000)
4. Everitt, B.S., Landau, S., Leese, M.: Cluster Analysis. Edward Arnold, London (2001)
5. Handl, J., Knowles, J., Dorigo, M.: Ant-Based Clustering: A Comparative Study of its relative performance with respect to k-means, average link and 1D-SOM, IRIDIA-Technical Report Series (2003)
6. Kohonen, T.: The self-organizing map. Neurocomputing 21, 1–6 (1998)
7. Kuo, R.J., Wang, H.S., Hu, T.-L., Chou, S.H.: Application of Ant K-Means on Clustering Analysis. Computers and Mathematics with Applications 50(10-12), 1709–1724 (2005)
8. Martens, D., De Backer, M., Haesen, R.: Classification with Ant Colony Optimization. IEEE Transactions on Evolutionary Computation 11(5), 651–665 (2007)
9. Milligan, G.W.: An Algorithm for generating Artificial Test Clusters. Psychometrika 50(1), 123–127 (1985)
10. Mitchell, T.: Machine Learning, 352 p. McGraw-Hill, New York (1997)
11. Sheng-Chai, C., Chih-Chieh, Y.: A Two-stage Clustering Method Combining Ant Colony SOM and K-means. Journal of Information Science and Engineering 24, 1445–1460 (2008)
12. Vesanto, J., Alhoniemi, E.: Clustering of the Self-Organizing Map. IEEE Transactions on Neural Networks 11(3), 586–600 (2000)
13. Yuqing, P., Xiangdan, H., Shang, L.: The k-means clustering algorithm based on density and ant colony. IEEE Intelligent Neural Networks and Signal Processing 1, 14–17 (2003)
14. Tan, P., Steinbach, M., Kumar, V.: Introduction to Data Mining, 769 p. Pearson, London (2006)