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. The aim of this method is not to find an optimal clustering for the data, but to obtain a view about the structure of data clusters. SOM is an Artificial Neural Network, which has one of its characteristics the nonlinear projection. AK is a meta-heuristic approach for solving hard combinatorial optimization problems based on Ant Colony Optimization (ACO). The SOMAK has a good performance when compared with some clustering techniques and reduces the computational time.

Keywords: SOM, ACO and Unsupervised Learning.

With the substantial reduction of cost data storage, 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 [1]. The techniques used in this paper were: K-means, SOM, SOM followed by the K-means (SOMK) and SOMAK. K-means is one of the simplest algorithms of non-supervised learning to solve the clustering problem. The aim is to divide the data set within k clusters fixed a priori [4]. AK [5] is a recently proposed meta-heuristic approach for solving hard combinatorial optimization problems named ACO [3].

The method proposed in this paper, SOMAK, can be seen in Fig. 1. SOMAK uses SOM net[1][2] 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. For the execution of the experiments were used: synthetic data (I, II, III), real data (IV, V), the method Monte Carlo, to check the efficiency of the clustering methods. The number of clusters and its centroids are obtained by SOM net and then uses AK to find the definite solution. Table 1 shows a comparison between SOMAK and SOMK to obtain a smaller number of clusters. It is important to mention in Table 1 that the fact of the SOMAK method increasing the number of clusters does not mean to say that bad, perhaps this increase may be necessary to have an improvement of entropy. We concluded that the experimental results are statistically independent according to the application of Test t and applied as well for the entropy [7] as for the computational time and with 5% of significance degree it showed that SOMAK is better than SOMK. The

[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/.

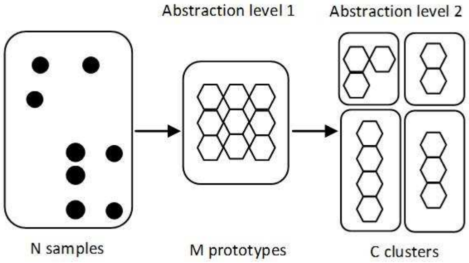

Figure 1: The first level of abstraction is obtained by using SOM. Algorithm SOMAK creates the second level of abstraction carrying out the cluster of M prototypes [6].

Table 1: 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

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. So, SOMAK is a robust method of clustering and it can be applied to a lot of different kinds 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.

References

[1] Everitt, B. S., Landau, S. and Leese, M.: Cluster Analysis. Edward Arnold. (2001)

[2] Kohonen, T.: The self-organizing map. Neurocomputing. **21** (1998) 1–6

[3] Dorigo, M. and Stützle, T.: The Ant Colony Optimization Metaheuristic: Algorithms, Applications, and Advances. Technical Report IRIDIA (2000)

[4] Mitchell, T.: Machine Learning. McGraw-Hill. (1997) 352p

[5] Kuo, R. J., Wang, H. S., Tung-Lai Hu and Chou, S. H.: Application of Ant K-Means on Clustering Analysis. Computers and Mathematics with Applications. **50** (2005) 1709–1724

[6] Vesanto, J. and Alhoniemi, E.: Clustering of the Self-Organizing Map. IEEE Transactions on Neural Networks. **11** (2000) 586–600

[7] Tan, P., Steinbach, M. and Kumar, V.: Introduction to Data Mining. Pearson. (2006)