AN IMPROVED MULTI-SOM ALGORITHM

Imen Khanchouch[1], Khaddouja Boujenfa[2] and Mohamed Limam[3]

[1]LARODEC ISG, University of Tunis
kh.imen88@gmail.com
[2]LARODEC ISG, University of Tunis
khadouja.Boujenfa@isg.rnu.tn
[3] LARODEC, ISG University of Tunis, Dhofar University, Oman
Mohamed.limam@isg.rnu.tn

ABSTRACT

This paper proposes a clustering algorithm based on the Self Organizing Map (SOM) method. To find the optimal number of clusters, our algorithm uses the Davies Bouldin index which has not been used previously in the multi-SOM. The proposed algorithm is compared to three clustering methods based on five databases. Results show that our algorithm is as performing as concurrent methods.

KEYWORDS

Clustering, SOM, multi-SOM, DB index.

1. INTRODUCTION

Clustering is an unsupervised learning technique aiming to obtain homogeneous partitions of objects while promoting the heterogeneity between partitions.In the literature there are many clustering categories such as hierarchical [13], partition-based [5], density-based [1] and neuronal networks (NN) [6].

Hierarchical methods aim to build a hierarchy of clusters with many levels. There are two types of hierarchical clustering approaches: the agglomerative methods (bottom-up) and the divisive methods (Top-down).Agglomerative methods start by many data objects taken as clusters and are successively joined two by two until obtaining a single partition containing all the objects. However, divisive methods begin with a sample of data as one cluster and successively get N divided clusters as objects. Hierarchical methods are time consuming in the presence of large amount of data. Consequently, the resulting dendrogram is very large and may include incorrect information.

Partitioning methods divide the data set into disjoint partitions where each partition represents a cluster. Clusters are formed to optimize an objective partitioning criterion, often called a similarity function, such as distance. Each cluster is represented by a centroid or a representative cluster. Partitioning methods suffer from the sensibility of initialization. Thus, inappropriate initialization may lead to bad results. However, they are faster than hierarchical methods.

Density-based clustering methods aim to discover clusters with different shapes. They are based on the assumption that regions with high density constitute clusters, which are separated by regions with low density. They are based on the concept of cloud of points with higher density where the neighborhoods of a point are defined by a threshold of distance or number of nearest neighbors.

DOI : 10.5121/ijnsa.2013.5414

NN are complex systems with interconnected neurons. Unlike hierarchical and partitioning clustering methods, NN can handle a large number of high dimensional data. Neural Gas is an artificial NN proposed by [11] which is based on feature vectors to find optimal representations for the input data. The algorithm's name refers to the dynamicity of the feature vectors during the adaptation process.

Based on competitive learning, SOM [6] is the most commonly used NN method. In the training process, the nodes compete to be the most similar to the input vector node. Euclidean distance is commonly used to measure distances between input vectors and output nodes' weights. The node with the minimum distance is the winner, also known as the Best Matching Unit (BMU). This latter, is a SOM unit having the closest weight to the current input vector after calculating the Euclidean distance from each existing weight vector to the chosen input record. Therefore, the neighbors of the BMU on the map are determined and adjusted. The main function of SOM is to map the input data from a high dimensional space to a lower dimensional one. It is appropriate for visualization of high-dimensional data allowing a reduction of data and its complexity. However, SOM map is insufficient to define the boundaries of each cluster since there is no clear separation of data items. Thus, extracting partitions from SOM grid is a crucial task. Also, SOM initializes the topology and the size of the grid where the choice of the size is very sensitive to the generalization of the method. Hence, we look for an extension of the multi-SOM to overcome these shortcomings and give the optimal number of clusters without any initialization.

This paper is structured as follows. Section 2 describes different clustering approaches. Section 3 details the multi-SOM approach and the proposed algorithm. Experimental results on real datasets are given in Section 4. Finally, a conclusion and some future works are given in Section 5.

2. THE MULTI-SOM APPROACH

The multi-SOM method was firstly introduced by [8] for scientific and technical information analysis specifically for patenting transgenic plant to improve the resistance of plants to pathogen agents. They proposed an extension of SOM, called multi-SOM, to introduce the notion of viewpoints into the information analysis with its multiple map visualization and dynamicity. A viewpoint is defined as a partition of the analyst reasoning. The objects in a partition could be homogenous or heterogeneous and not necessary similar. However objects in a cluster are similar and homogenous where a criterion of similarity is inevitably used. Each map in multi-SOM represents a viewpoint and the information in each map is represented by nodes (classes) and logical areas (group of classes).

[7] applied multi-SOM on an iconographic database. Iconographic is the collected representation illustrating a subject which can be an image or a document text. Then, multi-SOM model is applied in the domain of patent analysis in [10] and [9], where a patent is an official document conferring a right. The experiments use a database of one thousand patents about oil engineering technology and indicate the efficiency of viewpoint oriented analysis, where selected viewpoints correspond to: uses, advantages, patentees and titles subfields of patents.

[12] applied multi-SOM to a zoo data set from the UCI repository to illustrate the technique combining multiple SOMs which visualizes the different feature maps of the zoo data with color coded clusters superimposed. The multi-SOM algorithm supplies good map coverage with a minimal topological defects but it does not facilitate the integration of new data dimensions.

[4] applied the Multi-SOM algorithm for macrophage gene expression analysis. Their proposed algorithm overcomes some weaknesses of clustering methods which are the cluster number

estimation in partitioning methods and the delimitation of partitions from the output grid of SOM algorithm. The idea of [3] consists on obtaining compact and well separated clusters using an evaluation criterion namely Dynamic Validity Index (DVI). The DVI metric is derived from compactness and separation properties. Thus compactness and separation are two criteria to evaluate clustering quality and to select the optimal clustering layer. Compactness is assessed by the intra-distance variability which should be minimized and separation is assessed by the inter-distance between two clusters which should be maximized. The DVI metric is given by

$$DVI = \min_{k=1..K} \{IntraRatio(k) + \gamma InterRatio(k)\}$$

Where k denotes the number of activated nodes on the layer and γ is a modulating parameter.

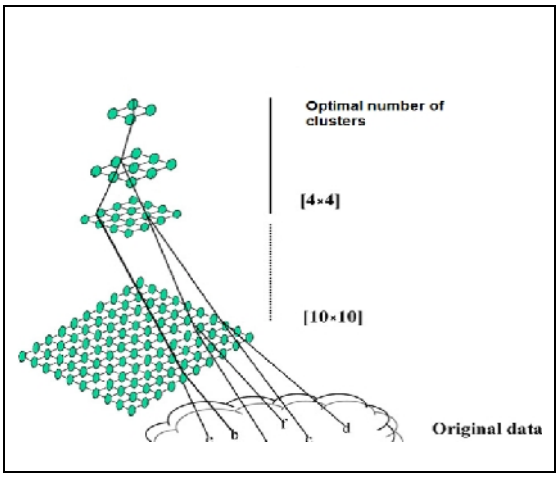

Figure 1. Architecture of the proposed multi-SOM algorithm

There are other evaluation criteria such as DVI. [2] used the DB index to measure cluster quality in Multi-Layer Growing SOM algorithm (GSOM) for expression data analysis. The DB index aims to define the compactness and how well separated are the clusters. GSOM is a model of NN algorithm which belongs to the hierarchical agglomerative clustering (HAC) algorithms. It is based on SOM approach but it starts with a minimum number of nodes which is usually four nodes and grows with a new node at every iteration.

We have chosen to use the DB index because it belongs to the internal criteria which is based on the compactness and separation of the clusters and well used in many works but not in the multi-SOM algorithm. The DB index is given by

$$DB = \frac{1}{c}\sum_{i=1}^{c} \max_{i \neq j} \left\{ \frac{d(X_i) + d(X_j)}{d(c_i, c_j)} \right\}$$

Where c defines the number of clusters, i and j denote the clusters, d(Xi) and d(Xj) are distances between all objects in clusters i and j to their respective cluster centroids, and d(ci, cj) is the distance between centroids. Smaller values of DB index indicate better clustering quality.

3. The proposed multi-SOM algorithm

The proposed algorithm (Algorithm1) aims to find the optimal clusters using the DB index as an evaluation criterion.

Given a dataset, the multi-SOM algorithm first uses the SOM approach to cluster the dataset and to generate the first SOM grid, namely SOM_1 and then the first SOM grid is iteratively clustered as shown in figure 1. The grid height (H_s) and the grid width (W_s) of SOM_1 are given by the user.

The multi-SOM algorithm uses the Batch map function proposed by [3] which is a version of SOM Kohonen algorithm but it is faster than SOM in the training process. Then, the SOM map is clustered iteratively from one level to another where the DB index is computed at each level. The size of the grid decreases at each level until the optimal number of clusters is reached.
In [4], the training process stops when one neuron is reached. However, our proposed algorithm stops when the DB index gets its minimum value. At this value the optimal number of clusters is given. Consequently, the proposed algorithm uses less computation time than the one proposed by [4].

The time complexity of DB is O (DB) = O(C) from the formula of DB index, where C is the number of clusters. This time complexity is less than the time complexity of the DVI, where O (DVI) = O (InterRatio + IntraRatio). This integrates many operations to compute the intra- and inter-distance. Thus, the computation of DVI values at each grid requires more memory space and time than that of the DB index.

The different steps of the algorithm are as follow:

s: the SOM layer current number
H_s : the SOMs grid height
W_s : the SOMs grid width
I_s : the input of SOMs
max_it: the maximum number of iterations for training SOM grids

Algorithm1: multi-SOM

Input: W_1, H_1, I_1, max_it
Output: Optimal cluster number, data partitions
Begin

• **Step1: Clustering data by SOM**
s = 1;
Batch SOM (W_1, H_1,I_1,max_it) ;
Compute DB index;
s = s+1;
• **Step2: Clustering of the SOM and cluster delimitation**

H_S=H_S-1;
W_S=W_S-1;
repeat
Batch SOM (W_s, H_s, I_s, max_it);
Compute DB index on each SOM grid;
s=s+1;
until(DB_s<DB_{s+1});
Return (Data partitions, Optimal cluster number);

End

4. EXPERIMENTAL RESULTS

This section gives the results of the proposed multi-SOM algorithm on five public datasets. Our algorithm is compared with a partitioning clustering method, namely k-means ([5]), a hierarchical method, namely BIRCH ([1]) and the algorithm proposed by [4]. The different data sets used in this work are extracted from the UCI machine learning repositories which are: iris, Pima Indians diabetes, wine, breast cancer and seeds.

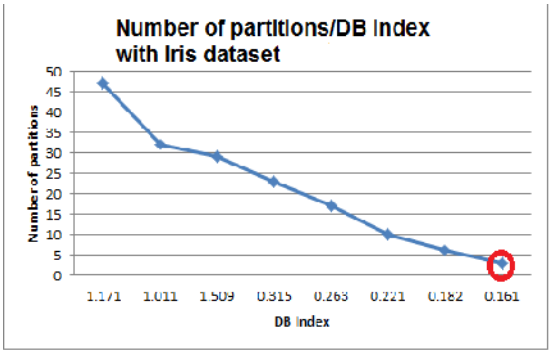

Figure 2. Variation of the number of partitions and the values of DB index with Iris

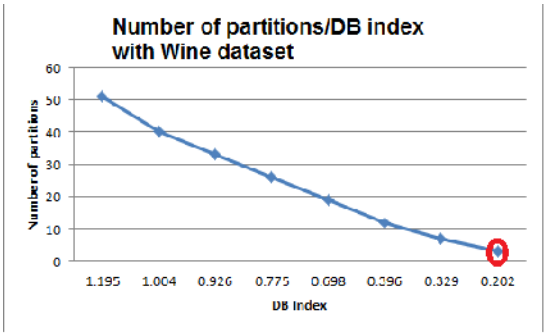

Figure 3. Variation of the number of partitions and the values of DB index with Wine

Figure 2 and 3 illustrate the variation of the DB index and the number of partitions for Iris and Wine datasets. We note that the DB index decreases gradually as the number of partitions decreases from one grid to another to obtain better clustering results until it reaches its lowest value of 0.161 and 0.202 for Iris and Wine datasets respectively. These values are relative to the optimal number of clusters which is 3 on Iris and Wine datasets

TABLE 1. Evaluation of the proposed multi-SOM algorithm

	Class number	k-means	Birch	Ghouila et al.(2008)	Multi-SOM
Iris	3	2	3	3	3
Pima	2	2	3	2	2
Breast cancer	2	2	2	2	2
Wine	3	5	4	3	3
Seeds	3	4	2	3	3

For all datasets, results given in Table (1) show that the proposed algorithm is as performing as that of [4]. On each dataset, the two algorithms succeeded in yielding the real number of clusters. However, k-means method determines the exact number of clusters only on Pima and breast datasets. On Iris, Breast and Seeds datasets, Birch algorithm generates the real number of clusters. With Wine dataset, the number of generated clusters given by the proposed multi-SOM algorithm is better than those given by k-means and Birch methods.

5. CONCLUSIONS

In this paper, a multi-SOM algorithm based on the DB index to determine the optimal number of clusters is proposed. In fact, the minimum value generated by DB index refers to the optimal cluster number. We have shown that our algorithm takes less iterations steps than that of [4]. Thus, the complexity of the proposed multi-SOM algorithm is better than the complexity of [4] algorithm. As a future work, we will investigate other validity indices and adapt our algorithm to fuzzy clustering.

REFERENCES

[1] Ester, M., Kriegel, H.-P., Sander, J., and Xu, X, (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In Proceedings of 2nd International Conference on KDD,Portland, Oregon, pp 226–231.

[2] Fonseka, A. and Alahakoon, D, (2010). Exploratory data analysis with multi-layer growing self-organizing maps. IEEE Xplore, pp 132–137.

[3] Fort J.C. Letrémy P. Cottrell M. (2002) , Advantages and Drawbacks of the Batch Kohonen Algorithm, Proc. ESANN 2002, Bruges, 2002, M.Verleysen Ed., Editions D Facto, Bruxelles p. 223-230.

[4] Ghouila, A., Yahia, S. B., Malouche, D., Jmel, H., Laouini, D., Z.Guerfali, F., and Abdelhak, S.(2008). Application of multisom clustering approach to macrophage gene expression analysis. Infection,Genetics and Evolution, Vol 9,pp 328–329.

[5] J.MacQueen (1967). Some methods for classification and analysis of multivariate observations. InProceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Vol 1, pp 281–289.

[6] Kohonen, T. (1981). Automatic formation of topological maps of patterns in a self-organising system. In Proceedings of the 2SCIA, Scand. Conference on Image Analysis, pp 214–220.

[7] Lamirel, J. C. (2002). Multisom: a multimap extension of the som model. Application to information discovery in an iconographic context.3:pp 1790–1795.

[8] Lamirel, J.-C., Polanco, X., and Francois, C. (2001). Using artificial neural networks for mapping of science and technology: A multi self-organizing maps approach. Scientometrics, 51:pp 267–292.

[9] Lamirel, J. C. and Shehabi, S. (2006). Multisom: a multimap extension of the som model. Application to information discovery in an iconographic context. IEEE CONFERENCE PUBLICATIONS, pp 42–54.

[10] Lamirel, J. C., Shehabi, S., Hoffmann, and Francois, C. (2003). Intelligent patent analysis through the use of a neural network: experiment of multi-viewpoint analysis with the multisom model. pp 7–23.

[11] Martinetz, T. and Schulten, K. (1991). A neural-gas network learns topologies. In Artificial Neural Networks.North-Holland, Amsterdam, pp 397–402.

[12] Smith, T. (2009). Adapting to increasing data availability using multi-layered self-organizing maps.IEEE Computer Society, pp 108–113.

[13] Zhang, T., Ramakrishna, R., and Livny, M. (1996). Birch: An efficient data clustering method for very large databases. pp 103–114.