A Comparative Study of Multi-SOM Algorithms for Determining the Optimal Number of Clusters

I. Khanchouch, M. Charrad, and M. Limam

Abstract—**The interpretation of the quality of clusters and the determination of the optimal number of clusters is still a crucial problem in clustering.**

We focus in this paper on multi-SOM clustering method which overcomes the problem of extracting the number of clusters from the SOM map through the use of a clustering validity index. We test the multi-SOM algorithm using real and artificial data sets with different evaluation criteria not used previously such as Davies Bouldin index, Dunn index and silhouette index. The multi-SOM algorithm is compared to k-means and Birch methods. Results show that it is more efficient than classical clustering methods.

Index Terms—**Clustering, SOM, multi-SOM, DVI, DB index, Dunn index, Silhouette index.**

I. Introduction

Clustering is considered as one of the most important tasks in data mining. It is a process of grouping similar objects or elements of data set into classes called clusters. The main idea of clustering is to partition a given set of data points into groups of similar objects where the notion of similarity is defined by a distance function. In the literature there are many clustering methods such as hierarchical, partition-based, density-based and neural networks (NN) and each one has its advantages and limits. We focus on neural networks especially Self Organizing Map (SOM) method. SOM, proposed by [1], it is the most widely used neural network method based on an unsupervised learning technique.

SOM method aims to reduce a high dimensional data to a low dimensional grid by mapping similar data elements together. This grid is used to visualize the whole data set.

However, SOM method suffers from the delimitation of clusters, since its main function is to visualize data in the form of a map and not to return a specified number of clusters.

That’s why a multi-SOM approach has been proposed by [2] to overcome this limit. To return the optimal number of clusters, [3] integrated a cluster validity index called Dynamic Validity Index (DVI) into the multi-SOM algorithm. Then, it is interesting to test this algorithm with other existing validity criteria.

In this paper, we study the existing clustering evaluation criteria and test multi-SOM with different validity indexes, then compare it with a partitioning and a hierarchical clustering method. We used R as a statistical tool to develop the multi-SOM algorithm.

The rest of this paper is structured as follows. Section II describes different clustering approaches. Section III details the multi-SOM approach and a literature review. Clustering evaluation criteria are given in Section IV. Finally, a conclusion and some future work are given in Section V.

II. Clustering Approaches

A. Hierarchical Methods

Hierarchical methods aim to build a hierarchy of clusters with many levels. There are two types of hierarchical clustering approaches namely agglomerative methods (bottom-up) and divisive methods (Top-down).

Divisive methods begin with a sample of data as one cluster and successively divide clusters as objects. However, the clustering in the agglomerative methods start by many data objects taken as clusters and are successively joined two by two until obtaining a single partition containing all objects.

The output of hierarchical methods is a tree structure called a dendrogram which is very large and may include incorrect information. Several hierarchical clustering methods have been proposed such as: CURE [4], BIRCH [5], and CHAMELEON [6].

B. Partitioning Methods

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 such as K-means [7], and PAM [8], suffer from the sensitivity of initialization. Thus, inappropriate initialization may lead to bad results.

C. Density-Based 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. Several density-based clustering methods have been proposed such as: DBSCAN [9] and OPTICS [10].

D. Neural Networks

Neural Networks are complex systems with high degree of interconnected neurons. Unlike the hierarchical and

Manuscript received November 12, 2014; revised March 5, 2015.

I. Khanchouch is with the LARODEC Laboratory and High Institute of Management, ISG Tunis, University of Tunis, Tunisia (tel.: +216 50 840 865; e-mail: imen.khanchouch@yahoo.fr).

M. Charrad is with University of Gabes, Tunisia (e-mail: malika.charrad@gmail.com).

M. Limam was with University of Tunis. He is now with Dhofar University, Oman (e-mail: Mohamed.limam@isg.rnu.tn).

DOI: 10.7763/IJFCC.2015.V4.384

partitioning clustering methods NN contains many nodes or artificial neurons so it can accept a large number of high dimensional data. Many neuronal clustering methods exist such as SOM and Neural Gas.

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). The 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. In fact, SOM output does not automatically give partitions, but its major function is to visualize a low dimensional map reduced from a high dimensional input data. 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 extend multi-SOM to overcome these shortcomings and give the optimal number of clusters without any initialization.

III. Multi-SOM Method

A. Definition

The multi-SOM is an unsupervised method introduced by [1]. Its main idea is the superposition and the communication between many SOM maps. The input data are firstly trained by SOM algorithm. Then, other levels of data are clustered iteratively based on the first SOM grid. Thus, the size of the maps decreases gradually since only a single neuron is obtained in the last layer. Each grid gathers similar elements into groups from the previous layer. It builds a hierarchy of SOM maps as shown in Fig. 1.

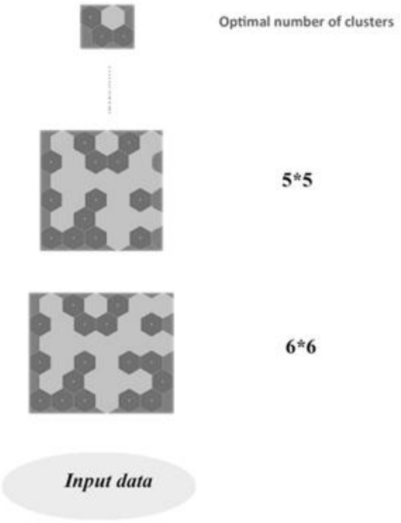

Fig. 1. Architecture of the multi-SOM approach.

B. Literature Review

The Multi-SOM method was firstly introduced by [1] for scientific and technical information analysis specifically for patenting transgenic plant to improve the resistance of the plants to pathogen agents.

Reference [1] proposed an extension of SOM called multi-SOM to introduce the notion of viewpoints into the information analysis with its multiple maps 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).

Reference [11] 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 [12] and [13], 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 of patents.

Reference [2] applied multi-SOM algorithm to 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 [2] consists on obtaining compact and well separated clusters using an evaluation criterion namely 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.

Reference [14] applied multi-SOM to real data sets to improve multi-SOM algorithm introduced by [2].

IV. Clustering Evaluation Criteria

The main problem in clustering is to determine the ideal number of clusters. Thus, cluster evaluation is usually used. In fact, many techniques and measures are used to test the quality of the clusters obtained as output data.

There are three categories of cluster evaluation namely: External validity measures, internal validity measures and relative validity measures.

- External criteria are based on the prior knowledge about data. They measure the similarity between clusters and a partition model. It is equivalent to have a labeled dataset. Many external criteria are cited in the literature like purity, entropy and F-measure.
- Relative criteria are based on the comparison of two different clusters or clustering results. The most well-known index is the SD index proposed by [15].
- Internal criteria are often based on compactness and

separation. That's why we focus on the internal validity indexes in this work to check the quality of clusters.

Compactness is assessed by the intra-distance variability which should be minimized. Separation is assessed by the inter-distance between two clusters which should be maximized.

Many internal criter*ia* exist such as: DB, Dunn, Silhouette, C, CH, DVI, etc. But, we focus on the following indexes:

- *Davies-Bouldin (DB)*

DB is proposed by [16] and given by:

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

where *c* is the number of clusters, *i* and *j* are the clusters, $d(X_i)$ and $d(X_j)$ are distances between all objects in clusters *i* and j to their respective cluster centroids, and $d(c_i, c_j)$ is the distance between these two centroids. Small values of DB index indicate good clustering quality.

- *Dunn Index (DI)*

DI is proposed by [17] and given by:

$$DI = \min_{1 \leq i \leq c}\left\{\min\left\{\frac{d\left(c_i, c_j\right)}{\max_{1 \leq k \leq c}\left(d\left(X_k\right)\right)}\right\}\right\} \tag{2}$$

where $d\left(c_i, c_j\right)$ denotes the distance between *ci* and *cj*

$d(X_k)$ represents the intra-cluster distance of the cluster X_k and c is the cluster number of the dataset.

Larger values of DI indicate better clustering quality.

- *Dynamic Validity Index (DVI)*

The DVI metric, introduced by [18], is derived from compactness and separation properties. Therefore, it considers both the intra-distance and the inter-distance which are defined as follow:

$$DVI = \min_{k=1..K}\left\{IntraRatio\left(k\right)+\gamma InterRatio\left(k\right)\right\} \tag{3}$$

$$IntraRatio(l) = \frac{Intra\left(l\right)}{MaxIntra\left(l\right)} \tag{4}$$

$$InterRatio(l) = \frac{Inter\left(l\right)}{MaxInter\left(l\right)} \tag{5}$$

where *l* is the layer of each grid and:

$$MaxIntra = \max_{l \in \{1..L\}}\left(Intra\left(l\right)\right) \tag{6}$$

$$Intra\left(l\right) = \frac{1}{N}\sum_{i=1}^{k_l}\sum_{x \in C_i}\left(\left\|x - z_i\right\|\right)^2 \tag{7}$$

$$MaxInter = \max_{l \in \{1..L\}}\left(Inter\left(l\right)\right) \tag{8}$$

$$Inter\left(l\right) = \frac{\max\left(\left\|z_i - z_j\right\|^2\right)}{\min_{i \neq j}\left(\left\|z_i - z_j\right\|^2\right)}\sum_{i=1}^{k_l}\left(\frac{1}{\sum_{j=1}^{kl}\left(\left\|z_i - z_j\right\|^2\right)}\right)$$

where *N* is the number of data samples Z_i and Z_j represent the reference vectors of nodes *i* and *j*.

The optimal number of clusters is determined by the minimal value of DVI in each level.

- *Silhouette*

This measure, introduced by [19], is defined by:

$$S = \frac{b\left(i\right)-a\left(i\right)}{\max\left\{a\left(i\right), b\left(i\right)\right\}} \tag{10}$$

where $a(i)$ is the average distance between the i^{th} sample and all of samples included in X_j, $b(j)$ is the minimum average distance between the i^{th} and all of the samples clustered in $X_k(k = 1... c; k \neq j)$.

Larger values of Silhouette index indicate better clustering quality.

V. Experiments

In this section, we carry out the evaluation of the multi-SOM algorithm on real data sets such as: Wine and Iris data sets with different clustering validity indexes as shown in Table I.

TABLE I: EVALUATION OF THE MULTI-SOM ALGORITHM ON REAL DATA SETS

Method	Correct Nb of clusters	DB	DVI	SIL	DUNN
Multi-SOM (Wine)	3	0.4	0.11	0.63	0.56
K-means (Wine)	5	0.49	0.49	0.55	0.53
Birch (Wine)	4	0.51	0.53	0.29	0.44
Multi-SOM (Iris)	3	0.55	0.32	0.38	0.64
K-means (Iris)	3	0.56	0.41	0.29	0.47
Birch (Iris)	5	0.71	0.48	0.18	0.25

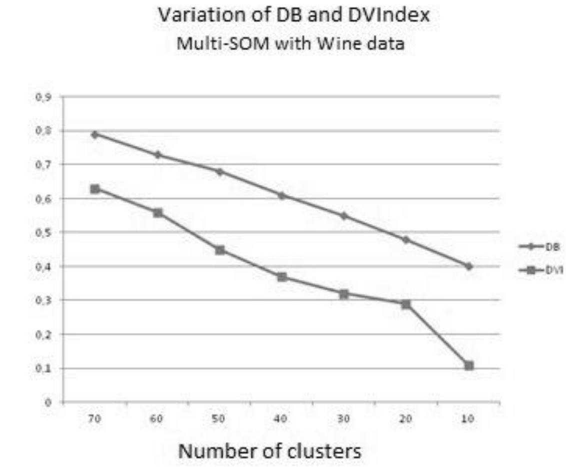

Fig. 2. Variation of DB and DV index.

Wine database is the result of a chemical analysis of wines derived from 3 different cultivars so this analysis determines the quantities of 13 constituents found in each of the three types of wines which are: Alcohol, Malic acid, Ash, Alcalinity of ash, Magnesium and Total phenols.

Iris is the most commonly used data base in the pattern recognition literature. It contains the characteristics of varieties of Iris plant. It contains 3 classes of 50 instances each one.

We have chosen 7 × 7 as dimension of the SOM map as the first SOM grid. Then, the number of clusters gradually decreases from a layer to another until we obtain the optimal number of clusters which is equal to 3.

TABLE II: EVALUATION OF THE MULTI-SOM ALGORITHM ON ARTIFICIAL DATA SETS

Method	Correct Nb of clusters	DB	SIL	DUNN
Circular Datasets				
Multi-SOM	2	0.41	0.38	0.47
K-means	2	0.58	0.22	0.39
Birch	2	0.69	0.55	0.52
Multi-SOM	3	0.5	0.26	0.66
K-means	2	0.44	0.22	0.47
Birch	2	0.44	0.21	0.41
Multi-SOM	5	0.4	0.36	0.488
K-means	4	0.42	0.32	0.45
Birch	3	0.61	0.28	0.33
Multi-SOM	8	0.33	0.28	0.44
K-means	7	0.51	0.16	0.41
Birch	6	0.53	0.15	0.38
Rectangular Datasets				
Multi-SOM	2	0.46	0.25	0.64
K-means	2	0.38	0.53	0.72
Birch	2	0.27	0.61	0.74
Multi-SOM	3	0.51	0.27	0.44
K-means	2	0.45	0.39	0.48
Birch	2	0.43	0.51	0.56
Multi-SOM	5	0.47	0.26	0.72
K-means	5	0.61	0.24	0.66
Birch	3	0.58	0.22	0.61
Multi-SOM	8	0.44	0.34	0.57
K-means	6	0.43	0.27	0.49
Birch	6	0.22	0.26	0.45
Elliptical Datasets				
Multi-SOM	2	0.52	0.25	0.42
K-means	2	0.46	0.22	0.37
Birch	2	0.43	0.2	0.26
Multi-SOM	3	0.47	0.28	0.54
K-means	2	0.45	0.21	0.41
Birch	3	0.34	0.13	0.22
Multi-SOM	5	0.502	0.37	0.49
K-means	4	0.39	0.35	0.45
Birch	3	0.4	0.33	0.39
Multi-SOM	8	0.507	0.21	0.73
K-means	7	0.4	0.12	0.71
Birch	6	0.38	0.11	0.67

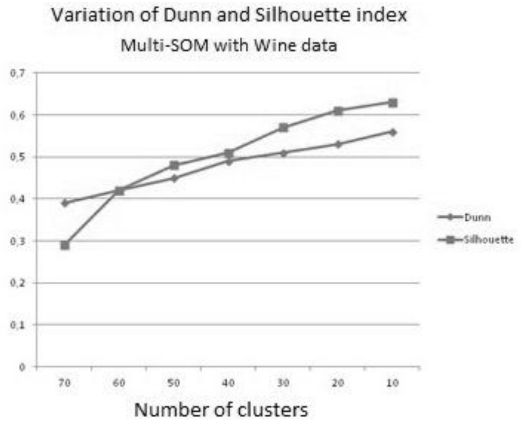

Fig. 3. Variation of Dunn and Silhouette index.

In Fig. 2, we notice that the optimal number of clusters is corresponding to the minimal value of DB and DBI index which are 0.4 and 0.11. However, in Fig. 3 the optimal number of clusters corresponds to the maximum value of Silhouette and Dunn index which are: 0.63 and 0.56.

Thus, we might simply conclude that DVI is more efficient than DB index and silhouette is more efficient than Dunn index.

We have also used 12 artificial data sets with different number of classes (2, 3, 5 and 8) and different shapes (circle, rectangle and ellipse) to test the different versions of multi-SOM algorithm as shown Table II.

To obtain these results, we developed a multi-SOM package using [20] R which is a statistical programming language.

Results show that the number of generated clusters given by the multi-SOM algorithm is usually better than those given by k-means and Birch methods.

VI. CONCLUSION

Classical clustering methods are developed by [21] to test 30 different validity indexes using R language.

Different clustering validity indexes are needed to assess the quality of clusters on each SOM grid. Compared with other classical clustering methods, multi-SOM is more efficient for the determination of the optimal number of clusters.

It could be applied to a wide variety of high dimensional data sets such as medical and banking data. As a future work we will apply multi-SOM algorithm for Market Segmentation.

REFERENCES

[1] T. Kohonen, “Automatic formation of topological maps of patterns in a self-organizing system,” in *Proc. the 2SCIA, Scand. Conference on Image Analysis*, 1981, pp. 214–220.

[2] J. C. Lamirel, “Using artificial neural networks for mapping of science and technology: A multi self-organizing maps approach,” *Scientometrics*, vol. 51, pp. 267–292, 2001.

[3] A. Ghouila, S. B. Yahia, D. Malouche, H. Jmel, D. Laouini, Z.Guerfali, and S. Abdelhak, “Application of multisom clustering approach to macrophage gene expression analysis,” *Infection, Genetics and Evolution*, vol. 9, pp. 328–329, 2008.

[4] S. Guha, R. Rastogi, and K. Shim, “Cure: An efficient data clustering method for very large databases,” in *Proc. ACM SIGMOD International Conference on Management of Data*, vol. 27, ACM Press, 1998, pp. 73–84.

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

[6] G. Karypis, E.-H. Han, and V. Kumar, “Chameleon: Hierarchical clustering using dynamic modeling,” *IEEE Xplore*, vol. 32, pp. 68–75, 1999.

[7] J. MacQueen, “Some methods for classification and analysis of multivariate observations,” in *Proc. the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, 1967, pp. 281–289.

[8] L. Kaufman and P. Rousseeuw, “Methods clustering by means of medoids,” *Statistical Data Analysis Based on the L1-Norm and Related*, pp. 405–417, 1987.

[9] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, “A density-based algorithm for discovering clusters in large spatial databases with noise,” in *Proc. 2nd International Conference on KDD*, Portland, Oregon, pp. 226–231, 1996.

[10] M. Ankerst, M. Breunig, H.-P. Kriegel, and J. Sander, “Optics: Ordering points to identify the clustering structure,” in *Proc. ACM SIGMOD International Conference on Management of Data*, June 1-3, Philadelphia, Pennsylvania, USA, vol. 28, 1999, ACM Press, pp. 49–60.

[11] J. C. Lamirel, “Multisom: A multimap extension of the som model,” *Application to Information Discovery in an Iconographic Context*, vol. 3, pp. 1790–1795, 2002.

[12] J. C. Lamirel and S. Shehabi, “Multisom: A multimap extension of the som model,” *Application to Information Discovery in an Iconographic Context*, IEEE Cobference Publications, pp. 42–54, 2006.

[13] J. C. Lamirel, S. S. Hoffmann, and C. Francois, “Intelligent patent analysis through the use of a neural network: Experiment of multi-viewpoint analysis with the multisom model,” pp. 7–23, 2003.

[14] I. Khanchouch, K. Boujenfa, and M. Limam, “An improved multi-SOM algorithm,” *International Journal of Network Security & Its Applications (IJNSA)*, vol. 5, no. 4, pp. 181-186, July 2013.

[15] M. Halkidi, M. Vazirgiannis, and Y. Batistakis, “Quality scheme assessment in the clustering process,” in *Proc. PKDD (Principles and Practice of Knowledge Discovery in Databases)*, Lyon, France, 2000.

[16] D. L. Davies and D. W. Bouldin, “A cluster separation measure,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 1, pp, 224-227, February 1979.

[17] J. C. Dunn, “A fuzzy relative of the isolate process and its use in detecting compact well-separated clusters,” *Cybernetics and Systems*, vol. 3, pp. 32–57, 1974.

[18] J. Shen, S. I. Chang, E. S., Lee, Y. Deng, and S. J. Brown, “Determination of cluster number in clustering microarray data,” *Applied Mathematics and Computation*, pp. 1172–1185.

[19] P. J. Rousseeuw, “Silhouettes: A graphical aid to the interpretation and validation of cluster analysis,” Computational and applied mathematics, vol. 20, pp. 53–65, November 1987.

[20] R C. Team. (2014). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*. [Online]. Available: URL http://www.R-project.org/

[21] M. Charrad, N. Ghazzali, V. Boiteau, and A. Niknafs, “NbClust: An r package for determining the relevant number of clusters in a data set,” *Journal of Statistical Software*, vol. 61, no. 6, pp. 1-36, 2014.

I. Khancouch is a PhD student at the High Institute of Management in Tunis and a member of LARODEC Laboratory. She received a bachelor of science (2010) in *Computer Science* and a MSc (2013) in statistics from High Institute of Management in Tunis.

M. Charrad is an assistant professor at Gabes University in Tunisia. She was a postdoctoral researcher in the Department of Mathematics and Statistics at Laval University in Quebec (2012-2013). She received a master of engineering in statistics (2003) and a MSc in computer science (2005) from the National School of Computer Science in Tunisia, and a PhD (2010) in computer science from the Conservatoire National des Arts et Méiers in France and La Manouba University in Tunisia. She is a member of RIADI Laboratory in Tunisia and a member of MSDMA team and CEDRIC Laboratory in CNAM, France. Her research interests are related to these topics data mining, web mining, and text mining, machine learning and social network analysis.

M. Limam is a professor of statistics at the University of Tunis. He received an MSc (1981) and PhD (1984) in statistics from Oregon State University, USA. He is the author of many research studies published in the Journal of the American Stat. Association, Machine Learning, Communications in Statistics, Quantitative Finance, Computer and Industrial Engineering, International Journal of Production Research, Quality and Reliability Engineering International, Chemometrics and Intelligent Laboratory Systems, Remote sensing letters, Bio Data Mining, Bio information. He is a founder of the Tunisian Association of Statistics and its Applications. Now, he is the vice president at Dhofar University in Oman.