Optimization of the Organized KOHONEN Map by a New Model of Preprocessing Phase and Application in Clustering

Es-safi Abdelatif
Scientific computing and Computer Sciences, Engineering Sciences
Modeling and Scientific Computing Laboratory
University of SIDI MOHAMED IBN ABDELAH
Faculty of sciences and techniques
FES, MOROCCO
essafiabdelatifi@hotmail.com

Harchli Fidae
Scientific Computing and Computer Sciences, Engineering Sciences
Modeling and Scientific Computing Laboratory
University of SIDI MOHAMED IBN ABDELAH
Faculty of sciences and techniques
FES, MOROCCO
Fidae.harchli@gmail.com

Ettaouil Mohamed
Scientific Computing and Computer Sciences, Engineering Sciences
Modeling and Scientific Computing Laboratory
University of SIDI MOHAMED IBN ABDELAH
Faculty of sciences and techniques
FES, MOROCCO
mohamedettaouil @yahoo.com

Abstract—**The needless of clustering is continually growing. This fact is due to the huge amount of information daily stored in the sites web. Those informations must be classified in order to facilitate their treatment. There are several methods to classify a set of data; those based on matching learning are more efficient. Among those systems, Kohonen algorithm is a useful system because of its famous proprieties. Some of them will be presented in this paper. Unfortunately, as other clustering algorithm, it suffers from the following problems: dependency of the result on the initialization phase which is performed randomly and the number of classes is unknown in the beginning. The overcome of these problems represents a great challenge in the clustering domain. In the present paper we expose an approach which allows performing suitably the initialization phase. This approach consists of conducting a pre-processing phase. In this latter we use a parameter r, we obtain an idea on distribution of examples. Then the initial weight vectors are chosen from the area which has a high density. This allows us to avoid an initialization with the isolated examples which decrease the performance of the system. Also we can determine approximately the number of classes. After measuring the quality of clustering obtained by Kohonen algorithm, we update the parameter r and we repeat the same process. This latter is arrested when we obtain a suitable quality of clustering. To show the performance of this approach, some experiments are conducted.**

General Terms—**Clustering, Standard Kohonen algorithm, Machine learning; Evaluation**

Index Terms—**Preprocessing phase, initial code-vectors, number of clusters, optimization**

I. INTRODUCTION

The clustering is an important tool in data mining. It consists of grouping a set of examples in some cluster. This plays a great importance in retrieval research (RI). It allows reducing the research space focusing on specific clusters. So the consuming times to access to the desired information is also reduced. This is more important than the case while size of the data set is great. Because of the enormous quantity of information available in tools communication, this is the case in almost actual situation. Adopting the matching learning techniques the performance of classifier are improved. According to the type of the learning used by the classifier, this latter can be supervised or unsupervised. In the first case, the class of the elements of the training set is determined. While in unsupervised case any information about the class is known. Among those algorithms the neural network Kohonen is an efficient algorithm: It has an important proprieties, among them we present the following:

- It is considered as a tool of reduction of dimensionality space, because each data will be associated to a dot in space with a low dimensionality. So it can be used as a technique of nonlinear projection and visualization of inputs as the PCA (Principal component analysis).
- It resist to noise.
- It is rapid in clustering of the new input.

As the other clustering system, the Kohonen algorithm suffers of the following drawbacks:

- Difficulty of determination of the suitable measure distance: In the numerical case (i.e. when the attributes tack numerical values), the standard measure is the Euclidian Manhattan and maximum distance measure. But in the case of categorical attributes the problem is more difficult.

doi:10.4304/jetwi.6.1.80-84

- Structure of database: Real life data may not always contains clearly identifiable cluster and contain several outliers.
- The determination of the number of clusters and the choice of the initial code-vectors are unknown: Opposite of the supervised case, the number of class in clustering system is unknown. Therefore the initial code-vectors are chosen randomly.

Outlier detection: The objects which are enormously dissimilar to other data are called outliers inputs. It is difficult to avoid such elements.

For more detail about the drawbacks of clustering system we refer the reader to [6].

In our work we focus on issue 2, 3 and 4 because identifying the number of class is a difficult task, a correct identification yield to good result. Otherwise a catastrophe can be expected. So it attracts a great interest in scientific research. Also the choice of initial clusters has a great effect in the result of a clustering system. Those systems choose the k initial cluster randomly. So it is likely that these classes are not suitable and can contain several outliers. This can produce empty classes or classes with few elements.

In this paper aiming to estimate a number of classes and suitable initial clusters, we conduct a study of data set. The precision increases during the process. The remaining of this paper is organized as follows: In section 2 we define the classical Kohonen algorithm. Our proposed method is introduced in section 3. During each iteration of the proposed algorithm, we evaluate the quality of the obtained clustering, so we reserve the section 4 to present the formula of evaluation. In section 5 we present the result obtained in experimental phase accompanied with discussion. In the last section we give a conclusion and some perspectives.

II Kohonen Algorithm

In the literature, Kohonen's SOM is a well-known and widely used algorithm in clustering techniques.

The learning process of Self organizing map is based on a competitive and unsupervised artificial neural network. It is a clustering algorithm that is used to map high-dimensional data into a low-dimensional representation space.

The underlying idea of this algorithm is to project a set of data called input space on an output space which has a low dimension (generally 3, 2 or 1). This is performed by establishing a correspondence between the two spaces. This latter must preserve the topology of data set i.e. if x is close to y; f(x) is also close to f(y). This idea is invented by scientist Tuevo Kohonen in 1975. The goal of this neural network is to model the brain activity. In order to realize this task he associates to each vector x another vector g(x) called the weight of the neuron f(x) and noted w (f(x) is noted by n, the set of neurons forms a gird in the output space). Each neuron has a neighborhood. The weight vector is initialized randomly and updated using a competitive and unsupervised learning. This adaptation process can be described as: For each presentation of an input x, the index i of the neuron nearest from x is determined, using the following formula:

$$d(x, w_i) = \min\left(d(x, w_j)\right), 1 \le j \prec N$$

Then the weight of the neuron ni and those of its neighborhood are updating as follows:

$$w_j(t) = w_j(t-1) + \eta(t)\alpha_{i,g}(t)\left(x - w_j(t-1)\right)$$

$w(i)$ represents a neuron in the neighborhood of the neuron n_i, η and α represents respectively the learning factor and neighborhood function which must decrease during the process. Indeed, the learning is performed on two phases; the first is called organization phase. This latter, because of a high rate learning and extended neighborhoods, allows deploying map were data are concentrated. After the learning rate, the size of the neighborhoods begins to decrease progressively, in order to conduct a folding of the map. In This phase the weight converge to some vectors which are approximation of the inputs. In the literature there exist different definitions of the parameters $\eta(t)$ and $\alpha_{i,g}(t)$. Among this latter we find the following:

$$\eta(t) = \begin{cases} \eta_0 - \left(\dfrac{\eta_0 - \eta_r}{r}\right)t & if\ t \prec \tau \\ \eta_r & otherwise \end{cases}$$

$\alpha_{i,g}(t) = \dfrac{1}{t}$ with t is the number of current iteration

To update the neighborhood of neurons some authors use the following formula:

$$v_{i,g}(t) = \exp\left(-\frac{d_{i,g}^2}{2\sigma^2(t)}\right)$$

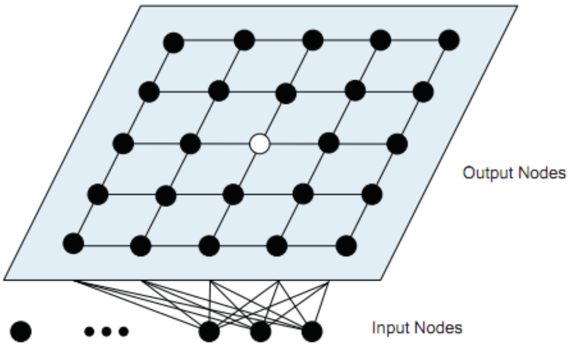

Fig 1: A self-organizing map with 25 neurons

III Proposed Method

In this section we present with more detail our approach. As is presented in previous algorithm the initialization of weight vectors is performed randomly. The number of initial cluster is also taken randomly. To overcome some of these difficulties, the searcher used different strategies. Some existing works hybrid supervised clustering and unsupervised one [2]. Some Latter works transform the problem of optimization of the architecture of Kohonen map to a problem of linear programmatic to improve the efficiency of Kohonen algorithm [4]. Other works are post clustering i.e. they choose the set of codes which is the partition and which provides the best partition from a number of alternative ones [1].

As is reported above a bad initialization leads to bad result. In this approach in order to initialize suitably weight vectors, we study the structure of input set. So we locate the areas where the density is great. Thereby the initial vector weights are taken from those. Hence we can remove the outliers and estimate the number of class. In this approach we propose two phases: the pretreatment phase and the clustering phase.

Practically, we begin by calculating the distance between all inputs which will be stored in a matrix $M = \left(a_{ij}\right)_{i,j}$ where $a_{ij} = d\left(x_i, x_j\right)$ and we search the maximum and the minimum in this matrix. Then we fix a real value $\varepsilon_1 = \frac{\max M - \min M}{n_{\max}}$. After, according to the parameter k data set E is divided on many subsets. Each one of these latter is formed as follow:

$$E_x = \left\{x_j / d\left(x, x_j\right) \prec \varepsilon_1\right\}$$

where x is an element chosen randomly from E. if the cardinality of E_x is equal to m (number of elements in data set) the first iteration in the pretreatment phase is achieved. Otherwise, another subset E_y is constructed by choosing randomly an element $y \in E \setminus E_x$ where $E_y = \left\{x_j / d\left(y, x_j\right) \prec \varepsilon_1\right\}$.

The same process will be applied to $E \setminus E_x \setminus E_y$ if $\left|E_x \cup E_y\right| \neq m$. It is certain that after a number of steps the process will stop and a number n of subsets is generated $E_1, E_2, ..., E_n$. We have conducted this phase in order to search the number of neurons in the map of Kohonen and the initial vector weights. So in the clustering phase, we turned the basic Kohonen algorithm with n neurons whose weights are the means of $E_i \ \forall i \in \left\{1, ..., n\right\}$. According to a measure function, the quality of the clustering is estimated in order to know the performance of the algorithm. In this stage we have generated the first iteration of our algorithm. The next iterations are performed by changing the value of the parameter ε_1 by $\varepsilon_k = \frac{\max M - \varepsilon_{k-1}}{n_{\max}}$ or $\varepsilon_k = \frac{\varepsilon_{k-1} - \min M}{n_{\max}}$ where $n_{\max}$ is a fixed number.

This approach is summarized in figure 1.

IV Assessment of Clustering

The evaluation of a system clustering represents a major challenge for the searchers. The quality of classifier is measured by calculating the homogeneity of classes that are produced and the separation between them. There exist several indexes which can be used to measure the quality of clustering. For more information we refer the reader to [3] [5].

In our work, we have chosen the average silhouette which is defined by the following formula:

$$s(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}}$$

This formula can be written as:

$$s(i) = \begin{cases} 1 - \frac{a(i)}{b(i)}, if\ a(i) < b(i) \\ 0, \quad if\ a(i) = b(i) \\ \frac{b(i)}{a(i)} - 1, \quad if\ a(i) > b(i) \end{cases}$$

Where: $a(i)$ is the average dissimilarity of xi with all other data within the same cluster and $b(i)$ is the lowest average dissimilarity to xi of any such cluster.

We report that: $a(i)$. More that $s(i)$ is close to 1 more than xi is correctly classified. In contrast, if $s(i)$ is close to -1 xi is appropriately clustered. Finally, if this value is about zero, xi is on the border of two natural clusters.

The definition of Silhouette index doesn't care into count the number of classes (n). Since in this work we attempt to look for a suitable number of classes, we developed a new index Ni.

$$N_i = \frac{(\text{Silhouette})^2}{\sqrt{\left(\frac{n}{150}\right)} + \text{Silhouette}}$$

This proposed index establishes a trade-off clustering quality and the number of classes.

V Experiments and Commentaries

In order to show the performance of our proposed method, two experiments are performed. In this context we used the data set IRIS. This latter is widely used is the quality clustering area. The inputs of this data set are ranked in three groups and they are characterized by four components.

In order to find a suitable number of classes we run several tests according to the parameter ε_k or a fixed number of iteration. In this experiment the best number of classes (s) corresponds to the best clustering quality according to the proposed criterion.

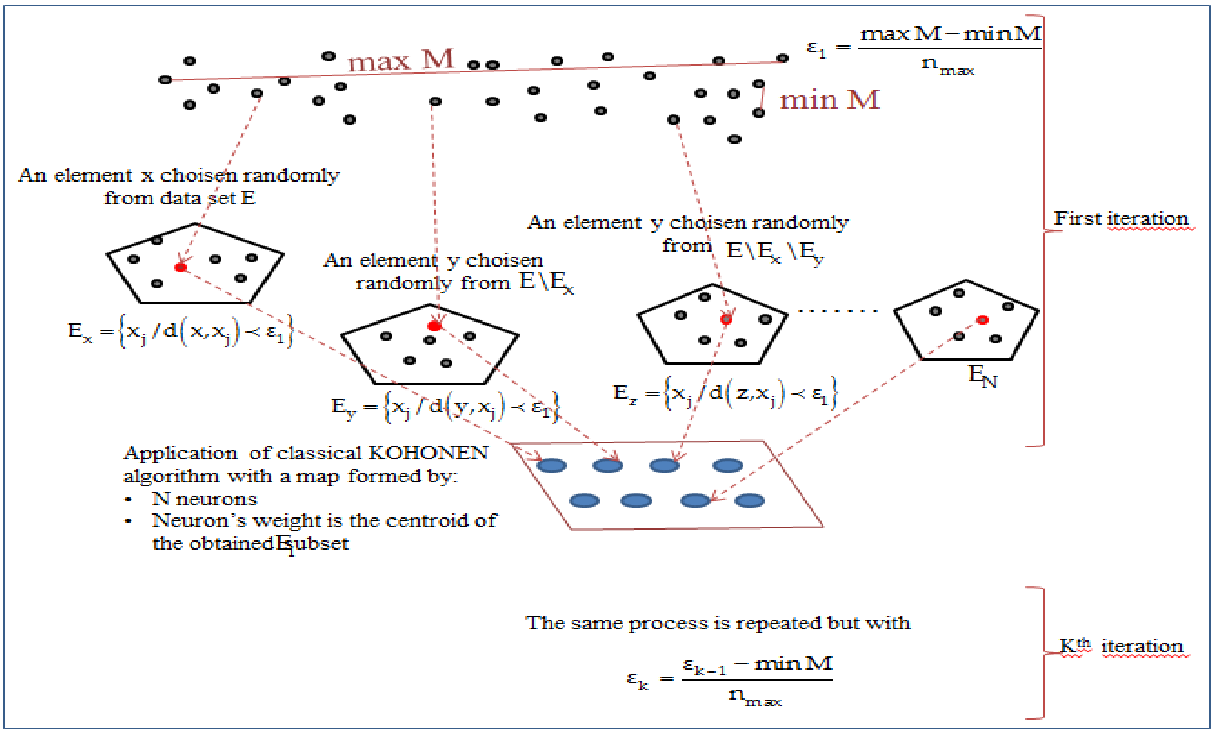
max M
$\varepsilon_1 = \frac{\max M - \min M}{n_{max}}$
min M
An element x choisen randomly from data set E
First iteration
An element y choisen randomly from $E \backslash E_x \backslash E_y$
An element y choisen randomly from $E \backslash E_x$
$E_x = \{x_j / d(x, x_j) \prec \varepsilon_1\}$
$E_y = \{x_j / d(y, x_j) \prec \varepsilon_1\}$
$E_z = \{x_j / d(z, x_j) \prec \varepsilon_1\}$
E_N
Application of classical KOHONEN algorithm with a map formed by:
N neurons
Neuron's weight is the centroid of the obtained E_i subset
The same process is repeated but with
$\varepsilon_k = \frac{\varepsilon_{k-1} - \min M}{n_{max}}$
K^{th} iteration

Fig 1: Summarization of the steps of the proposed algorithm

The following table reports the number of classes find by the system and clustering performance for different values of threshold ε_k.

TABLE 1.
OBTAINED NUMBER OF CLUSTERS IN THE PROCESS OF OPTIMIZATION

$\varepsilon_k = \frac{\max M - \varepsilon_{k-1}}{n_{max}}$		$\varepsilon_k = \frac{\varepsilon_{k-1} - \min M}{n_{max}}$	
Number of classes	**Proposed index**	**Number of classes**	**Proposed index**
2	0,25	2	0,25
3	0,28	3	0,77
3	0,77	5	0,77
2	0,25	15	0,69
3	0,28	14	0,7
3	0,77	16	0,69

In Table 1 we can see that whenever ε_k is close to maxM, the quality of the outputs become good but whenever ε_k is close to minM, inputs become dispersed and the system provides a large number of clusters. This experiment shows that with this method we achieved 0,77 of clustering performance with three clusters which is the desired number of classes. This satisfactory result proved the relevance of our approach.

To further investigate the relevance of the method we compare our results with those of classical Kohonen. In this latter we are forced to give a number of neurons and there weights. We built the map of classical Kohonen with s neurons (where s=3 is the result of the first experiment) and we affect a random initialization to these neurons. Results of this comparison are stored in Table II.

The proposed method is more powerful than conventional Kohonen algorithm and this is evident in Table II where we can see that the quality of result given by Kohonen classic is bad compared to what we found in the first experience. We can say that our approach gives a better result which is incomparable with the classical method.

TABLE 2.
COMPARISON BETWEEN THE RESULTS OF CLASSICAL KOHONEN AND THAT OF THE PROPOSED METHOD

Random initialization	Silhouette index	Proposed index
Choice 1	0,857	0,735
Choice 2	0,856	0, 734
Choice 3	0,853	0, 732
Choice 4	0,905	0, 782
Choice 5	0,906	0,784

Finally, those experiments show that the proposed method is able to identify suitable code vectors and give a number of classes equal to the real one. Therefore it gives a satisfactory clustering.

VI CONCLUSIVE REMARKS AND FUTURE WORKS

In this work a clustering method based on a clustering algorithm belonging to the Kohonen Self Organizing map family has been proposed. Despite its excellent properties whose some ones are represented above in this paper, the Kohonen algorithm is widely used in clustering area. Unfortunately, it suffers from some drawbacks whose main stems from its dependency on initialization phase. In this latter a number of cluster and code vectors are chosen randomly. So an inappropriate choice leads to a bad result. In this work, we proposed a method which aims to overcome this problem by choosing automatically suitable initial code vectors.

The analysis was performed on dataset IRIS. The experiments carried show that this method gives good result and satisfactory rate of clustering.

This is an encouraging result to try to integrate other parameters and criteria attempting to more improve the performance of clustering of this method, extend the analysis to datasets with more samples, implement this method on other databases, and apply this approach in other domain as text clustering and opinion meaning.

REFERENCES

[1] G. W. Milligan, and M. C. Cooper, "An Examination Of Procedures For Determining The Number Of Clusters in a Data Set," Psychometrica, vol. 50, No. 2, pp. 159-179, June 1985.

[2] K. Nigam, A. K. Mccallum, S. Thrun and T. Mitchell, “Text Classification from Labeled and Unlabeled Documents Using EM”, Machine Learning, vol. 39, pp.103-134, 2000.

[3] M. Ettaouil, A. Essafi, and F. Harchli, “Optimizing the architecture of Kohonen map and classification,” JOURNAL OF COMPUTING, vol. 4,pp. 31-38, July 2012.

[4] M. Ettaouil, Y. Gannou, K. Elmoutaouakil and M. Lazaar, “A new Architecture Optimization Model for the Network and Clustering,” Journal of Advanced Research in Computer Science, vol 3, no. 1, pp. 14-32, 2011.

[5] M. Qiu, S. Davis, and F. Ikem, “Evaluation of clustering Techniques in Data Mining Tools”, Issues Information systems, vol. 5, No. 1, 2004.

[6] P. Agarwal, M. Afshar Alam, and R. Biswas, " Issues, Challenges and Tools of Clustering Algorithms," IJCSI International journal of Computer Sciences Issues, vol. 8, Issue 3, No. 2, May 2011.