# K-MEANS AND SOM, A CONCURRENT VALIDATION SCHEME FOR DATA MINING

**MAURICIO SPERANDIO** **JORGE COELHO**
Federal University of Santa Catarina
Electrical Systems Planning Research Laboratory (LabPlan)
Florianópolis - Brazil
sperandio@labplan.ufsc.br, coelho@labplan.ufsc.br

***ABSTRACT***
We presents a concurrent validation scheme for clustering and data mining, were are exploit the knowledge discovery properties of the SOM (Self-Organizing Maps) for determine a good k-means clusters number estimation, and then visualize the k-means clusters over a trained map to compare and validate the result of both procedures. The SOM flexibility and visualization capacities are used to explore the results of a consolidated clustering algorithm like the k-means, and on the other hand, this method testify the clustering presented by the neural map. A real electric consumers database is analyzed for feature extraction.

## INTRODUCTION

Clustering methods are tools for exploratory data analysis in order to solve classification problems. The objective is to associate variable arguments (people, things, events, etc) in groups, or clusters, so that the similarity degree is big among members of a same group, and small among different groups. Then each group describes, in terms of the contained data, the class which their members belong, and abstract one or more private characteristics to denote them.

One of the most traditional clustering methods is the k-means, whose objective is to build a group of k groups starting from a mass of data, so that each unit belongs to just one group. It is a nonhierarchical method, so it don't make a linkage between the grouped data.

The Self-Organizing Maps (SOM) are a category of Artificial Neural Networks, characterized by an unsupervised "learning" process, in which the data are disposed over a plane (map) topologic organized according to similarity (Kohonen, 2001).

Many authors used to compare these two methods, pointing out the advantages and disadvantages of each one. However isn't common to use both together, to explore the best characteristics of each one. A work that used the idea of execute a competitive validation between the SOM and the k-means was presented in Coelho *et al*.(2002), but the methods didn't exchange information, and the objective was forecasting using a map trained with historical data.

In this paper is presented a concurrent validation between those two different methods, to certify that the discovered groups are really coherent in their formation.

**CONCURRENT VALIDATION**

First, it is required to define the variables that will take part in the database, what is an important aspect and can demand studies such as factorial analysis, to then make the procedures of clustering data with similar characteristics, knowledge discovery or subsequent classification.

In the proposed method, both procedures evaluate the same database, and the information exchange among the algorithms is firstly through the choice of the number of clusters to be formed by the k-means, based on the distance matrix of a trained map. So, the data are labeled with the k-means defined groups and printed in the winner neuron in the map. The validation is made verifying the k-means defined group that prevails in each neuron. As the map accomplish a topologic organization, where neighboring neurons have a high degree of similarity, configurations of k-means defined groups that are dispersed over the map are not accepted. In case of not being satisfied with the disagreement among the algorithms, one can change the number of clusters (k), or resize the map, looking for minimize the disagreements, the quantization and topologic errors.

Therefore, at the end of the concurrent validation process, the defined clusters are consolidated by two algorithms, and besides, the visualization advantage that the SOM offers, showing the groups organization over the map, facilitating the borders identification, as well as the component maps, that show the contribution of each variable for the map formation.

A flowchart of this procedure is presented in Figure 1, showing the relationship among each stage, in which happens some user intervention. First choice is the map size, then, verifying the U-Mat (Ultsch and Siemon, 1990), the Smoothed Data Histogram (Pampalk *et al.*, 2002) and the Component Maps should be determined a number of groups for the k-means; hence, the validation process occurs, and the user evaluates the errors report of crossing algorithms results to decide if he wants to change some parameter, if yes, which of them, the number k or the size of the map.

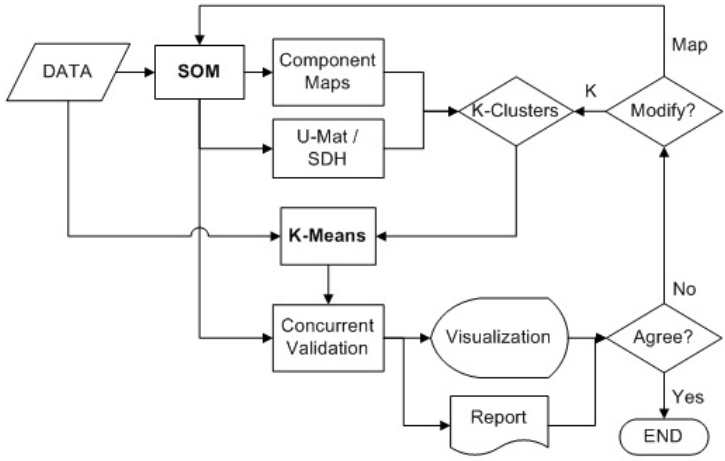


**Figure 1.** Concurrent Validation Flowchart (SOM x K-means).

The k-means and SOM algorithms were implemented in Matlab© by a group from the Neural Networks Research Center of the University of Helsinki, and are available on the internet (SOM Toolbox), documented by Vesanto *et al*. (1999). That toolbox is a library of codes that explore the flexibility of previous functions of Matlab, especially graphicals, for training and visualization of the maps, and other data that can be extract from this powerful method.

Therefore the whole process presented in the flowchart of Figure 1 runs in the Matlab environment, in which were developed the interface, the algorithm to minimize the k-means error, the concurrent validation process, and were also made adaptations to visualize the k-means results over the map.

To start the clustering process it is necessary to define a size for the self-organizing map, and that decision can affect the number of groups choice for the k-means. As the intention is just to accomplish clustering, an option is to begin with a small map, with approximately one neuron for each four entrance data, that tends to form concise groups. However, small maps may have a larger intra-cluster dispersion. Then, larger maps should be inspected after the first validation. Scattered maps take shape with more neurons for each data vector, and too many empty neurons can hinder neurons association to form larger groups, but reveals relationships among variables that are not evident in the database. Big maps are necessary when one want to use them for forecasting.

A good initial choice depends of the database characteristics. For instance, if the data tend to concentrate arround some value, a small map is enough to represent them. If they have a great dispersion in relation to the mean value, a big map characterizes them better. That initial estimation should be based in the descriptive analysis of the database variables that are being classified.

Concerning the process of concurrent validation, the two algorithms are independent, and the only information that is exchanged among them depends of the user decision when defining the k-means number of groups and the size of the self-organizing map. Even evaluating the errors table reported during the process, it should be aware that when increasing the number of groups, the intra-cluster dispersion is reduced, because less data will be allocated for each group, and when increasing the size of the map, the disagreement with the k-means decrease, because with more neurons there will be more space to allocate data, reducing border conflicts, that, except few cases, are the main reason of divergences among the two clustering methods.

Even with big maps, such exceptions seem to be result of a larger sensibility of the SOM to the data variability, while the k-means is fastened to a defined number of groups by the user.

The SOM can point out a new cluster partition when k-means defined groups are disconnected in the map topology, in other words, empty neurons appear between data defined as being of a same group by the other algorithm. Another form of increasing groups partition is verifying the component maps, that presents each variable density over the map, and therefore, it can be expected that exist a group with a great participation of one variable, other with a medium participation, other in that such variable is not representative, and etc.

**A REAL DATABASE CLUSTERING**

Following the presented methodology, a study was accomplished to determine the clustering of a database with variables that define the characteristics of the electric power system from 260 municipal districts of the Santa Catarina state, in the south of Brazil.

Four variables were chosen to describe the system configuration, City-Substation distance, Number of Feeders, Feeders Interchange Capacity and Distribution Grid Length.

The initial map was 8x7 neurons, and after the training the distance matrix results were evaluated to determine an initial k-means number of clusters. This is made with the help of the U-Mat and the SDH figures. The first one shows the distance between each neuron in gray scale, and also plots the neurons size proportional to the amount of data associated to them. The SDH do something similar to this, but with a contour plot that indicates clustering regions.

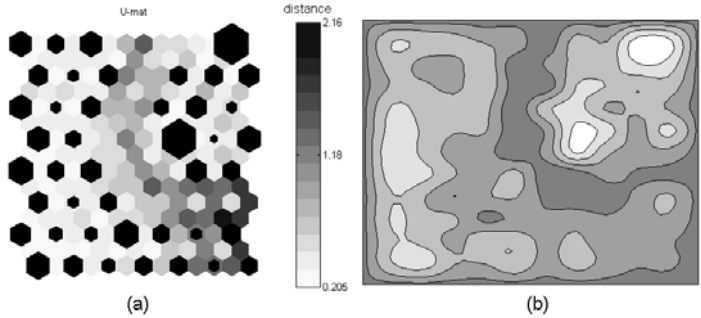


**Figure 2.**: (a) U-mat with density; (b) Smoothed Data Histograms (SDH).

Evaluating the graphs presented in Figure 2, quickly can be noted a division between two great groups, with an amount of small groups within. Figure 2(a) shows an uniform distribution of the data among the neurons, due to the fact that in this database discrepant values almost don't exist.

The initial number of clusters estimate to be formed by the k-means was 8, however along the process it indicated an increase to 10, above that, the new groups only contained 2 or 3 elements, and didn't bring significant information. The map size was chosen according to the number of clusters and the validation errors report, like showed in Table 1.

**Table 1.** Validation Errors Report for 10 clusters.

| Map Size | Quantization Error | Topographic Error | Disagree | Linked |
|---|---|---|---|---|
| 8x9 | 0,420 | 0,015 | 10 | no |
| 9x8 | 0,412 | 0,027 | 14 | no |
| 10x8 | 0,373 | 0,023 | 16 | yes |
| 8x10 | 0,393 | 0,012 | 10 | no |
| 10x9 | 0,353 | 0,019 | 9 | no |
| 10x10 | 0,341 | 0,019 | 6 | yes |
| 11x10 | 0,319 | 0,038 | 4 | yes |
| 12x11 | 0,294 | 0,012 | 7 | no |

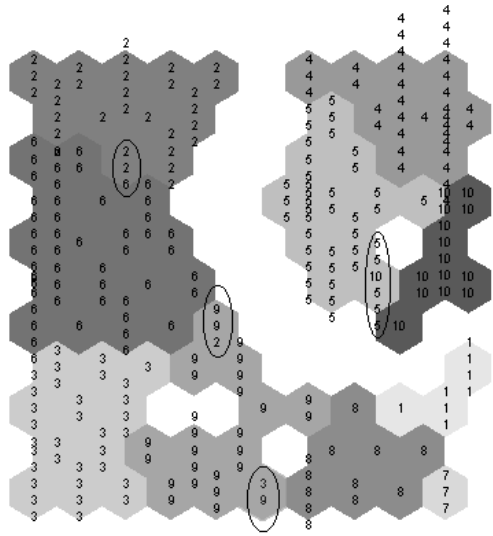

**Figure 3.** Best map with k-means group numbers.

Determined the map configuration (11x10 neurons, 10 clusters) we could observe the clusters distribution over it. Figure 3 shows the k-means clusters numbers over their winners neurons, the gray shades delimit clusters of neurons after the validation, and the ellipses emphasizes the four discordants data, where two are border conflicts (6-2 and 10-5) and the others are really a divergence among the algorithms (2-9 and 3-9). That divergent data are closer to the SOM validated cluster than to the k-means determined number.

The next figure shows the component maps that form Figure 3, and aid to understand each cluster characteristics. Each map represent only one variable, and the shades are its participation in each neuron.

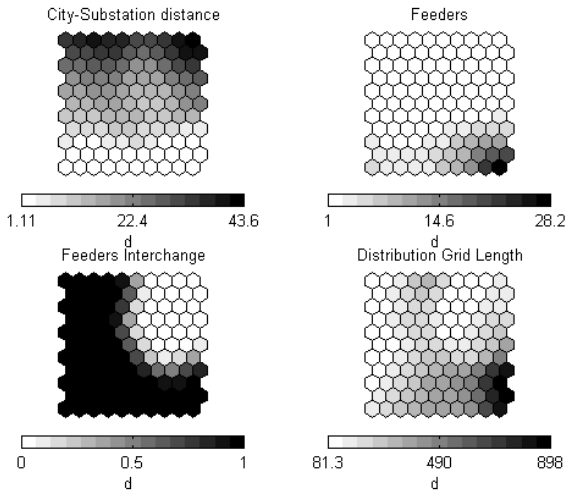


**Figure 4.** Four variables component maps .

The Feeders Interchange variable explains the division of the two cited larger groups in Figure 2, with or without this capability. There we can see that in the map area without Feeders Interchange (white), are also neurons with larger City-Substantion Distance, and they still have only 1 Feeder. For the brazilian case, that is explained by the small Distribution Grid Length, in other words, that neurons are associated with small municipal districts with little load. Consequently those districts should be the ones that have the worst supply continuity indexes.

After the electric system clustering, a cluster ranking was made based on their mean supply continuity indexes. The same cluster analysis was applied to electricity demand variables, and hence, the crossing of ranked clusters of demand and system attributes indicated municipal districts where the utility must apply system reinforcements to have adequate supply quality, or where the quality goal set by the regulatory body is to tight.

## CONCLUSION

The presented validation scheme between the k-means statistical method and the Self-Organizing Maps (SOM) showed to be a flexible and reliable tool, because the best characteristics of both algorithms are explored .

Allying the rigid and recognized k-means clustering, with the capacity of knowledge discovery and visualization of the SOM, permit to look for the best groups configuration to represent the attributes of the contained data, allowing borders identification, as well as the most susceptible elements to change group. This is an important characteristic, because it makes possible a quickly reclassification, without running the algorithm again, just renumbering the wanted border neuron.

This scheme also shows that the SOM is a good clustering tool, that used with the appropriated knowledge could lead to excellent results. This work allowed an utility to argue with the brazilian energy regulatory body for suitables quality goals and adequate investments refund.

## REFERENCES


Pampalk, E., Rauber, A., Merkl, D., 2002: Using Smoothed Data Histograms for Cluster Visualization in Self-Organizing Maps, *Proc. of the International Conference on Artificial Neural Networks*,.Springer Lecture Notes in Computer Science, Madrid, Spain.

Ultsch, A., Siemon, H.P, 1990: Kohonen's Self-Organizing Feature Maps for Exploratory Data Analisys, *Proc. Intern. Neural Networks*, pp: 305 – 308, Paris, France.

Kohonen, T., 2001, *Self-Organizing Maps*; Springer-Verlag, 3ª ed.

Vesanto, J., Himberg, J., Alhoniemi, E., Parhankangas, J., 1999: Self-Organizing Map in Matlab: the SOM Toolbox, *In Proc. of the Matlab DSP Conference*, pp: 35–40, Espoo, Finland.

Coelho, J., Gauche, E., Queiroz, H., Nassar, S.M., Ricardo,V.W., Lima, M., 2002: Influence of Weather Variables in Continuity Levels of Electrical Power Supply – An Analysis through Artificial Neural Networks, *Symposium of Specialists in Electric Operational Expansion Planning* (*SEPOPE)*, Curitiba, Brazil.