Self-organizing Maps as Substitutes for K-Means Clustering

Fernando Bação[1], Victor Lobo[1,2], and Marco Painho[1]

[1] ISEGI/UNL, Campus de Campolide, 1070-312 LISBOA, Portugal
bacao@isegi.unl.pt
[2] Portuguese Naval Academy, Alfeite, 2810-001 ALMADA, Portugal
vlobo@isegi.unl.pt

Abstract. One of the most widely used clustering techniques used in GISc problems is the k-means algorithm. One of the most important issues in the correct use of k-means is the initialization procedure that ultimately determines which part of the solution space will be searched. In this paper we briefly review different initialization procedures, and propose Kohonen's Self-Organizing Maps as the most convenient method, given the proper training parameters. Furthermore, we show that in the final stages of its training procedure the Self-Organizing Map algorithms is rigorously the same as the k-means algorithm. Thus we propose the use of Self-Organizing Maps as possible substitutes for the more classical k-means clustering algorithms.

1 Introduction

The widespread use of computers and Geographical Information Systems (GIS) made available a huge volume of digital geo-referenced data (Batty and Longley 1996). This growth in the amount of data made multivariate data analysis techniques a central problem in Geographical Information Science (GISc). Amongst these techniques, cluster analysis (Jain, Murty et al. 1999) is one of the most used, and it is usually done using the popular k-means algorithm. Research on geodemographics (Openshaw, Blake et al. 1995; Birkin and Clarke 1998; Feng and Flowerdew 1998; Openshaw and Wymer 1994), urban research (Plane and Rogerson 1994; Han, Kamber et al. 2001), identification of deprived areas (Fahmy, Gordon et al. 2002), and social services provision (Birkin, Clarke et al. 1999) are examples of the relevance that clustering algorithms have within today's GISc research.

There have been a number of tests comparing SOM's with k-means (Balakrishnan, Cooper et al.1994; Openshaw and Openshaw 1997; Waller, Kaiser et al. 1998). Conclusions seem to be ambivalent as different authors point to different conclusions, and no definitive results have emerged. Some authors (Flexer 1999; Balakrishnan, Cooper et al. 1994; Waller, Kaiser et al. 1998) suggest that SOM performs equal or worst than statistical approaches, while other authors conclude the opposite (Openshaw and Openshaw 1997; Openshaw, Blake et al. 1995).

The main objective of this paper is to analyze the performance of the SOM and k-means in clustering problems, and evaluate them under specific conditions. We

V.S. Sunderam et al. (Eds.): ICCS 2005, LNCS 3516, pp. 476 – 483, 2005.

review both algorithms and then compare their performance on specific problems, using two synthetic datasets, and three real-world datasets.

2 K-Means Algorithm and Its Initialization

The k-means algorithm is widely known and used so only a brief outline is presented (for a thorough review see (Kaufman and Rousseeuw 1990; Fukunaga 1990; Duda, Hart et al. 2001)). K-means is an iterative procedure, to place cluster centers, which quickly converges to a local minimum of its objective function (Bradley and Fayyad 1998; Kanungo, Mount et al. 2002). This objective function is the sum of the squared Euclidean distance (L2) between each data point and its nearest cluster center (Selim and Ismail 1984; Bradley and Fayyad 1998). This is also known as "square-error distortion" (Jain and Dubes 1988). It has been shown that k-means is basically a gradient algorithm (Selim and Ismail 1984; Bottou and Bengio 1995) which justifies the convergence properties of the algorithm. he original online algorithm (MacQueen 1967) is as follows:

```
Let         k   be the predefined number of centroids
            n   be the number of training patterns
            X   be the set of training patterns x1, x2,..xn
            P   be the set of k initial centroids μ1, μ2,… μk taken from X
            η   be the learning rate, initialized to a value in ]0,1[
1   Repeat
2           For i=1 to n
3               Find centroid μj∈P that is closer to xi
4               Update μj by adding to it Δμj = η(xi - μj)
5           Decrease η
6           Until η reaches 0
```

There are a large number of variants of the k-means algorithm. In this study we use the generalized Loyd's algorithm (Duda, Hart et al. 2001), which yields the same results as the algorithm above (Bottou and Bengio 1995). The popularity of this variant in statistical analysis is due to its simplicity and flexibility. It does not, however, specify how the initial centroids should be selected. Due to the gradient nature of the algorithms, these initial centroids have a decisive effect on which areas of the solution space can be searched. In all but the simplest cases the solution space contains many local optima to which the k-means algorithm may converge. To guarantee that a good solution will be found, multiple initializations of the algorithms are usually tested, and only the best final solution is kept.

By far the most common initialization, called "Forgy Approach" (Peña, Lozano et al. 1999), consists on randomly selecting *k* of available data patterns as centroids. This method has as main advantage it's simplicity: the selection requires no prior knowledge or computational effort, and multiple initializations will usually cover rather well the solution space. This is the initialization procedure used by default by software packages such as SAS Enterprise Miner, Matlab, and Clementine.

Instead of choosing *k* samples, we may divide the dataset into *k* subsets, and then use the centroids of these sets as seeds. We wil call this the "random selection method" (Peña, Lozano et al. 1999). It is similar to calculating the centroid of the whole dataset and then obtaining *k* perturbations of this point (Thiesson, Meek et al. 1999). More random selection based algorithms have been proposed (Kaufman and Rousseeuw 1990; Cano, Cordón et al. 2002; MacQueen 1967) each with specific strengths and weaknesses. The order by which the initial seeds are presented may influence the final outcome, so re-ordering techniques have been used in (Fisher, Xu et al. 1992) and (Roure and Talavera 1998). Sensitivity to outliers is another important problem that can be minimized by repeating random selection. Various methods can be used to repeat selection and clustering on smaller datasets, such as proposed by by (Bradley and Fayyad 1998).

Genetic algorithms are a well established technique to "guide randomness", and thus can be used to generate successive random selections. This approach is followed by (Peña, Lozano et al. 1999). Several attempts have been made to avoid randomness in the selection of seeds by using deterministic density estimation methods, and selecting the points of higher density as seeds. Such is the approach followed in (Bradley and Fayyad 1998), (Fukunaga 1990). Another family of initialization methods comes from heuristics that use the distance between candidate seeds as a guide for their selection (Katsavounidis, Jay Kuo et al. 1994;Al-Daoud and Roberts 1994;Tou and González 1974).

Hierarchical clustering algorithms are widely used and can produce meaningful clusters of data, but they usually do not minimize the objective function of the k-means algorithm. They may however be used to obtain a good approximation that can be used as seed for it's initialization. This approach was proposed in (Fisher 1987), and is used under different forms by (Higgs, Bemis et al. 1997; Snarey, Terrett et al. 1997;Meila and Heckerman 2001). The major drawback of these types of initializations is that they require a lot of computational effort.

Comparing results with all these initialization techniques is a Herculean task, and since simple random selection (or "Forgy Approach") is the most common and simple, we will use it in the comparisons with SOM.

3 Self-organizing Maps and Their Use in Obtaining K-Clusters

Although the term "Self-Organizing Map" could be applied to a number of different approaches, we shall use it as a synonym of Kohonen's Self Organizing Map (Kohonen 1982;Kohonen 2001), or SOM for short, also known as Kohonen Neural Networks.

The basic idea of a SOM is to map the data patterns onto a *n*-dimensional grid of neurons or units. That grid forms what is known as the output space, as opposed to the input space where the data patterns are. This mapping tries to preserve topological relations, i.e., patterns that are close in the input space will be mapped to units that are close in the output space, and vice-versa. So as to allow an easy visualization, the

output space is usually 1 or 2 dimensional. The basic SOM training algorithm can be described as follows:

```
Let  X  be the set of n training patterns x1, x2,..xn
     W  be a p×q grid of units wij where i and j are their
        coordinates on that grid
     α  be the learning rate, assuming values in ]0,1[, initialized
        to a given initial learning rate
     r  be the radius of the neighborhood function h(wij,wmn,r),
        initialized to a given initial radius
1  Repeat
2    For k=1 to n
3      For all wij∈W, calculate dij = || xk - wij ||
4      Select the unit that minimizes dij as the winner wwinner
5      Upadate each unit wij∈W: wij = wij + α h(wwinner,wij,r) || xk - wij ||
6    Decrease the value of α and r
7  Until α reaches 0
```

The neighborhood function h is usually a function that decreases with the distance (in the output space) to the winning unit, and is responsible for the interactions between different units. During training, the radius of this function will usually decrease, so that each unit will become more isolated from the effects of its neighbors. It is important to note that many implementations of SOM decrease this radius to 1, meaning that even in the final stages of training each unit will have an effect on its nearest neighbors, while other implementations allow this parameter to decrease to zero.

SOMs can be used in many different ways, even within clustering tasks (Bação, Lobo et al. 2005). In this paper we will assume that each SOM unit is a cluster center, and thus a k-unit SOM will perform a task similar to k-means. It must be noted that SOM and k-means algorithms are rigorously identical when the radius of the neighborhood function in the SOM equals zero (Bodt, Verleysen et al. 1997). In this case the update only occurs in the winning unit just as happens in k-means (step 4).

4 Experimental Setting

4.1 Datasets Used

The data used in the tests is composed of 4 basic datasets, two synthetic and two real-world. The real-world datasets used are the well known iris dataset (Fisher 1936) and sonar dataset (Sejnowski and Gorman 1988). The iris dataset has 150 observations with 4 attributes and 3 classes, while the sonar dataset has 208 observations with 60 attributes and 2 classes. Two synthetic datasets were created. The first dataset, DS1, comprises 400 observations in two-dimensions with 4 clusters. Each of these clusters has 100 observations with a Gaussian distribution around a fixed center. The variance of these Gaussians was gradually increased during our experiments. The second data set, DS2, consists of 750 observations with 5 clusters with Gaussian distributions defined in a 16 dimensional space.

4.2 Robustness Assessment Measures

In order to access the performance of the two methods a set of three measurements was used. The first one is the quadratic error i.e., the sum of the squared distances of each point to the centroid of its cluster. This error is divided by the total dispersion of each cluster so as to obtain a relative measure. This measure is particularly relevant as it is the objective function of the k-means algorithm. Additionally, the standard deviation of the mean quantization error is calculated in order to evaluate the stability of the results found in the different trials. The second measure used to evaluate the clustering is the mean classification error. This measure is only valid in the case of classification problems and is the number of observations attributed to a cluster where they do not belong. Finally, a structural measurement is used in order to understand if the structural coherence of the groups is preserved by the clustering method. This measure is obtained by attributing to each cluster center a label based on the labels of the observations which belong to its Voronoi polygon. If more than one centroid receives a given label (and thus at least one of the labels is not attributed) then the partition is considered to be structurally damaged.

5 Results

Each dataset was processed 100 times by each algorithm, and the results presented in table 1 constitute counts or means. Table 1 presents a summary of the most relevant results. A general analysis of table 1 shows a tendency for SOM to outperform k-means. The mean quadratic error over all the datasets used is always smaller in the case of the SOM, although in some cases the difference is not sufficiently large to allow conclusions. The standard deviation of the quadratic error is quite enlightening showing smaller variations in the performance of the SOM algorithms. The class error indicator reveals a behavior similar to the mean quadratic error. Finally, the structural error is quite explicit making the case that SOM robustness is superior to k-means.

Looking closer at the results in different datasets, there is only one data set in which k-means is not affected by structural errors. The reason for this is related with the configuration of the solution space. In the sonar dataset the starting positions of the k-means algorithm are less relevant than in the other 3 datasets.

Table 1. – Comparison of SOM and k-means on different datasets, using the average quadratic error. its standard deviation, average classification error, and average structural error, over 100 independent initializations

Dataset	Method	Quadratic error	Std(Qerr)	ClassErr	Struct Err
IRIS	SOM	86.67	0.33	9.22	0
	k-means	91.35	25.76	15.23	18
SONAR	SOM	280.80	0.10	45.12	0
	k-means	280.98	3.18	45.34	0
DS1	SOM	9651.46	470.36	1.01	0
	k-means	11341.49	2320.27	12.77	58
DS2	SOM	27116.40	21.60	7.40	0
	k-means	27807.97	763.22	15.51	49

The real-world dataset refers to enumeration districts (ED) of the Lisbon Metropolitan Area and includes 3968 ED's which are characterized based on 65 variables, from the Portuguese census of 2001. Exploratory analysis of this dataset using large size SOMs and U-Matrices suggests that we should consider 6 clusters within this dataset. To find the exact locations and members of these 6 clusters we applied a batch k-means algorithm to this data, and compared the results with those obtained with a 6x1 SOM. In both cases we repeated the experiment 100 times with random initializations. The quadratic error obtained with k-means was 3543 ± 23 with a minimum of 3528, whereas with SOM we obtained 3533 ± 6 with a minimum of 3529. These results show that the best clustering obtained with each method is practically the same, but on average SOM outperforms k-means and has far less variation in it's results.

6 Conclusions

The first and most important conclusion that can be drawn from this study is that SOM is less prone to local optima than k-means. During our tests it is quite evident that the search space is better explored by SOM. This is due to the effect of the neighborhood parameter which forces units to move according to each other in the early stages of the process. This characteristic can be seen as an "annealing schedule" which provides an early exploration of the search space (Bodt, Cottrell et al. 1999). On the other hand, k-means gradient orientation forces a premature convergence which, depending on the initialization, may frequently yield local optimum solutions.

It is important to note that there are certain conditions that must be observed in order to render robust performances from SOM. First it is important to start the process using a high learning rate and neighborhood radius, and progressively reduce both parameters to zero. SOM's dimensionality is also an issue, as our tests indicate that 1-dimensional SOM will outperform 2-dimensional matrices. This can be explained by the fact that the "tension" exerted in each unit by the neighboring units is much higher in the case of the matrix configuration. This tension limits the plasticity of the SOM to adapt to the particular distribution of the dataset. Clearly, when using a small number of units it is easier to adapt a line than a matrix.

These results support Openshaw's claim which points to the superiority of SOM when dealing with problems having multiple optima. Basically, SOM offers the opportunity for an early exploration of the search space, and as the process continues it gradually narrows the search. By the end of the search process (providing the neighborhood radius decreases to zero) the SOM is exactly the same as k-means, which allows for a minimization of the distances between the observations and the cluster centers.

References

Al-Daoud, M. and S. Roberts (1994). New Methods for the Initialisation of Clusters. Leeds, University of Leeds: 14.

Bação, F., V. Lobo, M. Painho (2005). "The Self-Organizing Map, Geo-SOM, and relevant variants for GeoSciences." Computers & Geosciences, Vol. 31, Elsevier, pp. 155-163.

Balakrishnan, P. V., M. C. Cooper, V.S. Jacob, P.A. Lewis (1994). "A study of the classification capabilities of neural networks using unsupervised learning: a comparison with k-means clustering." Psychometrika 59(4): 509-525.

Batty, M. and P. Longley (1996). Analytical GIS: The Future. Spatial Analysis: Modelling in a GIS Environment. P. Longley and M. Batty. Cambridge, Geoinformation International: 345-352.

Birkin, M. and G. Clarke (1998). "GIS, geodemographics and spatial modeling in the UK financial service industry." Journal of Housing Research 9: 87-111.

Birkin, M., G. Clarke, M. Clarke (1999). GIS for Business and Service Planning. Geographical Information Systems. M. Goodchild, P. Longley, D. Maguire and D. Rhind. Cambridge, Geoinformation.

Bishop, C. M. (1995). Neural Networks for Pattern Recognition, Oxford University Press.

Bodt, E. d., M. Cottrell, M. Verleysen (1999). Using the Kohonen Algorithm for Quick Initialization of Simple Competitive Learning Algorithms. ESANN'1999, Bruges.

Bodt, E. d., M. Verleysen, M. Cottrell (1997). Kohonen Maps versus Vector Quantization for Data Analysis. ESANN'1997, Bruges.

Bottou, L. and Y. Bengio (1995). Convergence Properties of the K-Means Algorithms. Advances in Neural Information Processing System. Cambridge, MA, MIT Press. 7 G: 585-592.

Bradley, P. and U. Fayyad (1998). Refining initial points for K-means clustering. International Conference on Machine Learning (ICML-98).

Cano, J. R., O. Cordón, F. Herrera, L. Sánchez (2002). "A Greedy Randomized Adaptive Search Procedure Applied to the Clustering Problem as an Initialization Process Using K-Means as a Local Search Procedure." International Journal of Intelligent and Fuzzy Systems 12: 235-242.

Duda, R. O., P. E. Hart, D. Stork (2001). Pattern Classification, Wiley-Interscience.

Fahmy, E., D. Gordon, S. Cemlyn (2002). Poverty and Neighbourhood Renewal in West Cornwall. Social Policy Association Annual Conference, Nottingham, UK.

Feng, Z. and R. Flowerdew (1998). Fuzzy geodemographics: a contribution from fuzzy clustering methods. Innovations in GIS 5. S. Carver. London, Taylor & Francis: 119-127.

Fisher, D. H. (1987). "Knowledge Acquisition Via Incremental Conceptual Clustering." Machine Learning 2: 139--172.

Fisher, D. H., L. Xu, N. Zard (1992). Ordering effects in clustering. Ninth International Conference on Machine Learning, San Mateo, CA.

Fisher, R. A. (1936). "The use of Multiple Measurements in Taxonomic Problems." Annals of Eugenics VII(II): 179-188.

Flexer, A. (1999). On the use of self-organizing maps for clustering and visualization. Principles of Data Mining and Knowledge Discovery. Z. J.M. and R. J., Springer. 1704: 80-88.

Fukunaga, K. (1990). Introduction to statistical patterns recognition, Academic Press Inc.

Han, J., M. Kamber, A. Tung (2001). Spatial clustering methods in data mining. Geographic Data Mining and Knowledge Discovery. H. Miller and J. Han. Londor, Taylor & Fancis: 188-217.

Higgs, R. E., K. G. Bemis, I. Watson, J. Wikel (1997). "Experimental Designs for Selecting Molecules from Large Chemical Databases." Journal of Chemical Information and Computer Sciences 37(5): 861-870.

Jain, A. K. and R. C. Dubes (1988). Algorithms for clustering data, Prentice Hall.

Jain, A. K., M. N. Murty, P. Flynn (1999). "Data Clustering: A review." ACM Computing Surveys 31(3): 264-323.

Kanungo, T., D. M. Mount, N. Netanyahu, C. Piatko, R. Silverman, A. Wu (2002). "An efficient k-means clustering algorithm: analysis and implementation." IEEE Transactions on Pattern Analysis and Machine Intelligence 24(7): 881-892.

Katsavounidis, I., C.-C. Jay Kuo, Z. Zhang (1994). "A new initialization technique for generalized Lloyd iteration." IEEE Signal Processing Letters 1(10): 144 - 146.

Kaufman, L. and P. J. Rousseeuw (1990). Finding groups in data : an introduction to cluster analysis. New York, John Wiley & Sons.

Kohonen, T. (1982). Clustering, Taxonomy, and Topological Maps of Patterns. Proceedings of the 6th International Conference on Pattern Recognition.

Kohonen, T. (2001). Self-Organizing Maps. Berlin-Heidelberg, Springer.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observation. 5th Berkeley Symposium on Mathematical Statistics and Probability, University of California Press.

Meila, M. and D. Heckerman (2001). "An Experimental Comparison of Several Clustering and Initialization Methods." Machine Learning 42: 9-29.

Openshaw, S., M. Blake, C. Wymer (1995). Using neurocomputing methods to classify Britain's residential areas. Innovations in GIS. P. Fisher, Taylor and Francis. 2: 97-111.

Openshaw, S. and C. Openshaw (1997). Artificial intelligence in geography. Chichester, John Wiley & Sons.

Openshaw, S. and C. Wymer (1994). Classifying and regionalizing census data. Census Users Handbook. S. Openshaw. Cambridge, UK, Geo Information International: 239-270.

Peña, J. M., J. A. Lozano, P. Larrañaga (1999). "An empirical comparison of four initialization methods for the k-means algorithm." Pattern recognition letters 20: 1027-1040.

Plane, D. A. and P. A. Rogerson (1994). The Geographical Analysis of Population: With Applications to Planning and Business. New York, John Wiley & Sons.

Roure, J. and L. Talavera (1998). Robust incremental clustering with bad instance orderings: a new strategy. IBERAMIA 98 - Sixth Iberoamerican Conference on Artificial Intelligence, Lisbon, Springer Verlag.

Sejnowski, T. J. and P. Gorman (1988). "Learned Classification of Sonar Targets Using a Massively Parallel Network." IEEE Transactions on Acoustics, Speech, and Signal Processing 36(7): 1135 -1140.

Selim, S. Z. and M. A. Ismail (1984). "k-means type algorithms: a generalized convergence theorem and characterization of local optimality." IEEE Trans. Pattern Analysis and Machine Intelligence 6: 81-87.

Snarey, M., N. K. Terrett, P. Willett, D. Wilton (1997). "Comparison of algorithms for dissimilarity-based compound selection." Journal of Molecular Graphics and Modelling 15(6): 372-385.

Thiesson, B., C. Meek, D. Chickering, D. Heckerman (1999). Computationally Efficient Methods for Selecting Among Mixtures of Graphical Models. Bayesian Statistics. J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith. Oxford, UK, Oxford University Press. 6.

Tou, J. and R. González (1974). Pattern Recognition Principals. Reading, MA, Addison Wesley Publishing Company.

Waller, N. G., H. A. Kaiser, J. Illian, M. Manry (1998). "A comparison of the classification capabilities of the 1-dimensional Kohonen neural network with two partitioning and three hierarchical cluster analysis algorithms." Psychometrika 63(1): 5-22.