Comparison between Mahalanobis distance and Kullback-Leibler divergence in clustering analysis

ALLAN DE MEDEIROS MARTINS
ADRIÃO DUARTE DÓRIA NETO
JORGE DANTAS DE MELO
Department of Computation and Automation
Federal University of Rio Grande do Norte
Lagoa Nova, Cep.: 59072-970 Natal, RN
BRASIL

Abstract: - This paper shows a comparison between two clustering algorithms that use divergence measures to aid the clustering task. Both algorithms take a N-dimensional data set and uses competitive neural networks to separate them into isotropic clusters. Those clusters are then grouped based on a divergence measure. In this paper we compare that procedure using two different divergence measures, the Mahalanobis distance and the Kullback-Leibler divergence. The paper shows some results of clustering on both algorithms and make a few comments about the choice of the free parameters in both situations.

Key-Words: - clustering, neural networks, pattern classification, data mining, dada analysis

1 Introduction

The clustering technique is widely used in many areas such as data mining, image segmentation, pattern recognition, statistical data analysis and others. In those areas, the clustering is an important task that is part of the whole process. The main objective of a good clustering algorithm is to separate classes (or clusters) distributed arbitrarily in the data space of the data set, in a non-supervised way. In this paper we compare two algorithms used for clustering. Both are based on a new approach first proposed in [1]. Some others techniques to archive this task have been developed and are based on several heuristics. The most simple way of clustering or classification is the use of vector quantization techniques like k-means[2] or pure competitive neural networks[3] to find centers that represents the regions, segmenting the data set into isotropic clusters. The similarity measure used in most of these techniques is the Euclidian distance between the point and the center of its class. More elaborate techniques, using Kohonen(SOM) maps[4] or Fuzzy k-means[5] have been developed. In the SOM algorithm, the centers that represent the classes have a topologic organization (each center have neighbors) organized in to a MAP. The Fuzzy k-means technique uses the concept of pertinence function of the point to each center. That pertinence indicates how strong that center belongs to a center, it goes from 0 (don't belongs) to 1 (belongs). Some modifications of the usual SOM algorithm based in the segmentation of the output map was also been developed[6][7] and give good results, but need complex computation tasks. In that technique, the output map in the SOM algorithm is used to compute the distance (in fact the Euclidian distance) from each center to their neighbors, forming a distance matrix that is segmented (using image processing techniques). In most used techniques, the number of classes that exists in the data set must be given *a priori*. In some cases this information is not available, so it's important to develop algorithms that perform the automatic classification without this information. Another problem in clustering techniques is the complexity of the spatial distribution of the data set. The two approaches compared here uses a simple competitive neural network and a linking (or grouping) heuristic. That heuristic group similar clusters using a divergence measure as metric of dissimilarity between them. One of the algorithms uses the Kullback-Leibler divergence[3] and the other uses the Mahalanobis distance[8]. Both measures incorporate the spatial statistics of the data, giving us a good measure of the distribution of the points, making possible the algorithm to be used to classify very complex data sets. The only two *a priori* information about the data set given to the algorithm is the number of auxiliary centers and a threshold dissimilarity.

The main advantage of the methods compared here and the others clustering methods is that in most methods (we can say, k-means, competitive neural networks, SOM-Maps and SOM-Map based methods, and others), we must inform *a priori* the number of classes present in the data set. Another

advantage is the computational cost. In some methods like showed in [6] and [7], the computational cost is high, because they use complex tools for segmenting the output map in the Kohonen algorithm. The algorithms compared here use a simple competitive neural network to initially cluster the data and make comparisons between measures in the initial clusters. That has a low computational cost and after the clustering, the classification of new input data is made just measuring the probability to each cluster as in a Bayes classifier[12].

The paper is organized as follows; in the section 2 we will describe the approach used in the algorithms; in the section 3 we will describe the Kullback-Leibler and the Mahalanobis distance. The section 4 shows some clustering results using both approaches and in the section 5 we present some conclusions and comments about the comparison.

2 Clustering Approach

The algorithms compared here uses the approach described in[1] and [9]. In that approach, the entire data set is segmented into small isotropic clusters using some vector quantization technique. After that, the small clusters (called here *regions*) are grouped based on a measure of divergence between them. The divergency is measure to each pair of small clusters and they are grouped or not based on a threshold value of that divergence. The vector quantization technique used in this work was pure competitive neural networks. The figure 1 shows a data set after quantization.

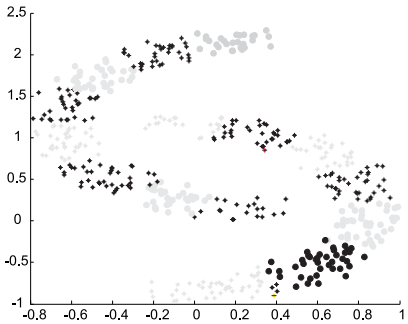

Fig 1: Quantized data set.

As we can see, the points in the data set are quantized to one among several centers, grouping the points in small regions. The number of centers is chosen arbitrarily. In this clustering technique the result depends on the choice of number of centers, but as we can see in the results, its robust and we obtain the same result to a wide rang of values.

3 Divergence Metrics

To make a decision whether goup or not group two regions, we must use some criteria. In [1] that criteria is based on a threshold in a measure of the Mahalanobis distance between two centers of two regions. In this section, we will explain how to use the Mahalanobis distance and the Kullback-Leibler divergence to measure the divergence between the regions.

2.1 Mahalanobis distance

The Mahalanobis metrics is a similarity measure that consider the spatial statistics of the points where the measure is been made[10]. The distance between two points $\mathbf{p}_1$ and $\mathbf{p}_2$ inside a space where the covariance matrix of the distribution of the probability that represents the spatial statistics is $\mathbf{C}$, is given by:

$$d_m(\mathbf{p}_1, \mathbf{p}_2, \mathbf{C})^2 = (\mathbf{p}_1 - \mathbf{p}_2)\mathbf{C}^{-1}(\mathbf{p}_1 - \mathbf{p}_2)^t \qquad (1)$$

The figure 2 shows the effect of the spatial probability in the distance of two points. In that figure, even the distance d_1 looks greater (in terms of Euclidian distance) it is smaller than the distance d_2. That difference exists because the covariance matrix of the spatial distribution gives a greater weight to the distances that doesn't been in the direction of the distribution.

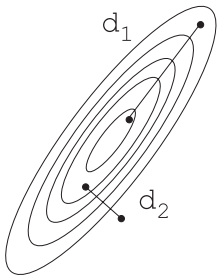

Fig. 2: Distance measurement in a space with non-uniform probability distribution

The Mahalanobis distance is important in classification problems because the information about spatial distribution of the points if incorporated in the metrics. The spatial distribution can be represented by the statistics of the data set where we wish to measure, in some way, the distance between two points. We can consider the Euclidian distance, a special case of the Mahalanobis distance, where the data would be uniformly distributed. That case corresponds a distribution where the covariance matrix is a

diagonal matrix. In that sense, the Mahalanobis metrics become a general case of distance measurements and suitable to be used in classification problems. In this work, we measure the divergence between two clusters, by the Mahalanobis distance to their centers considering $\mathbf{C}$ as the covariance matrix of the a region formed by the points that belong to both clusters. This approach leads to values that are as low as the clusters are "aligned" or in other words, have same statistics.

2.1 Kullback-Leibler divergence

The dissimilarity between two clusters is a number that measure how distinct they are. The greater the number, more distinct the clusters are. There are several ways to measure dissimilarity between two clusters[8]. The most common is the distance of their centers. This measure is good when we have isotropic clusters where we have points equally distributed in all directions. In most clustering cases, we have non-isotropic clusters that makes the Euclidian distance not good to measure dissimilarity. To measure the dissimilarity between clusters in a clustering algorithm the aim is to detect the separation between them. In the figure 3 we can see two situations, where clusters are separated and where are not.

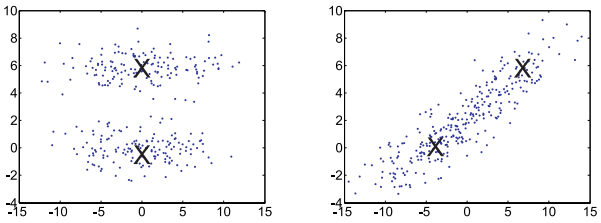

Fig. 3: Regions in two different situations

As we can see, the Euclidian distance between both situations are almost the same, but we have in one case a similar cluster and a dissimilar cluster in the other. One solution to better measure the dissimilarity was given in [1]. In that paper the authors uses the Mahalanobis distance as dissimilarity distance. In this paper we present another dissimilarity measure to use with a cluster algorithm, the Kullback-Leibler divergence. The Kullback-Leibler is a measure based on the relative entropy of two probability density functions[3][11]. The Kullback-Leibler divergence between two probabilities density p and q for a given random variable $\mathbf{X}$ is given by

$$D_{p\|q} = \sum_{\mathbf{x}\in\mathbf{X}} p(\mathbf{x})\log\left(\frac{p(\mathbf{x})}{q(\mathbf{x})}\right) \qquad (2)$$

This divergence measures how different they are on representing the data $\mathbf{X}$. We can use that property to measure if two clusters are similar or not in a statistical point of view. That will be explained in the next sections.

The estimative of $p(\mathbf{x})$ and $q(\mathbf{x})$ can be calculated by any method, in this paper we use Parzen windows[12] to obtain $p(\mathbf{x})$ and $q(\mathbf{x})$. The equation (3) shows how to estimate a probability density function from a given set of data.

$$p(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{V_n}\varphi\left(\frac{\mathbf{x}-\mathbf{x}_i}{h_n}\right) \qquad (3)$$

where $\varphi(\mathbf{x})$ is the window function and n is the number of points. V_n and h_n are the volume and edge length of a hypercube where the function will be evaluated.

3 Clustering results

The algorithm with both metrics was tested in a three-dimensional data set (showed in figure 4). Both algorithms archive the correct clustering, each one with an appropriate threshold. The figure 5 shows the result of clustering and the parameters used to archive that clustering.

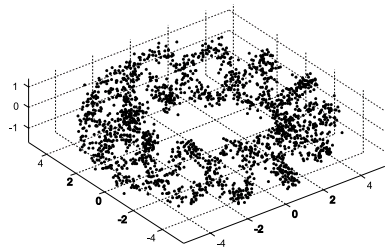

Fig. 4: Original data set used to test the algorithms (two springs interlaced)

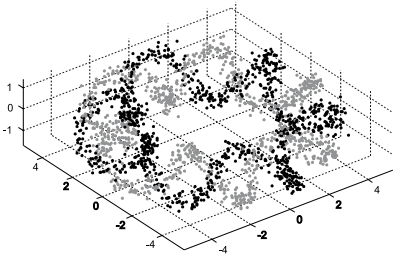

Fig. 5: Classified rings data set. For Mahalanobis metrics we get $d_t = 65$ and for Kullback-Leibler divergence we get $d_t = -15$. Both use 80 auxiliares centers (initial regions).

The figures 6 and 7 shows the results in terms of number of classes found as function of threshold distance d_t to both approaches. As we can see, using the Mahalanobis distance we have a wide range of threshold values that gives the same number of classes (2 classes in that case). That number is exactly the number of classes present in the data set. That makes possible the correct choice of the threshold distance by analyzing the results. With that procedure, we can make the choice of the threshold distance an automatic task.

With the kullback-leibler divergence, we can't use the results to help the choice of the threshold because the range of values of threshold that give the correct number of classes (2 in that case) is not significant.

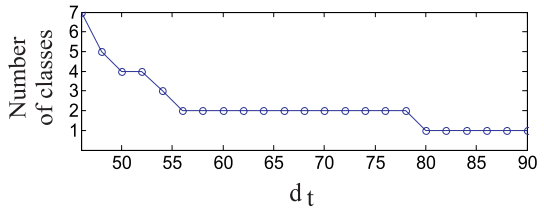

Fig. 6: Results of number of classes versus threshold distance using Mahalanobis distance.

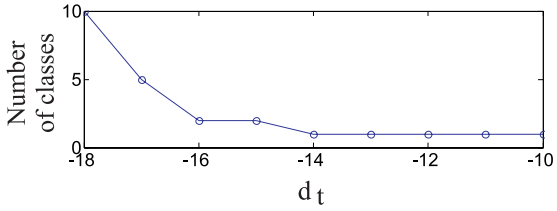

Fig. 7: Results of number of classes versus threshold distance using Kullback-Leibler divergence

In order to test the robustness of the algorithms we have tested it with different number of initial regions in the vector quantization stage. The figures 8 and 9 shows the results (also in terms of number of classes) as function of number of number of initial regions. The figures shows that, for larges number of initial regions, we have a stable situation in both cases. Because of that result, we can conclude that the choice of number of initial regions is robust for larges numbers of initial regions. Of course, if the number is made too much large the effect of local statistics is lost and the algorithm fails. To know how large that number can be, we can look for results like showed in the figures 8 and 9.

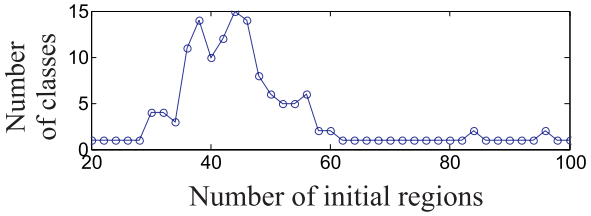

Fig 8: Results of number of classes versus number of initial regions using Mahalanobis distance.

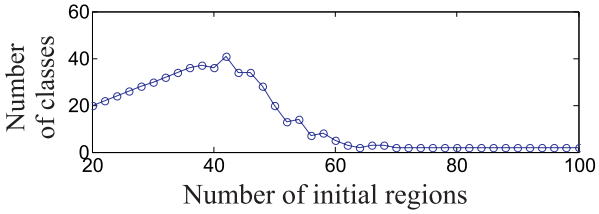

Fig 9: Results of number of classes versus number of initial regions using Kullback-Laibler divergence.

4 Conclusion

We have compared the use of Kullback-Leibler divergence and Mahalanobis distance in clustering task. The results showed that both metrics are capable to archive the correct clustering using the correct threshold and number of initial centers.

The approach used still depends on the correct choice of the threshold d_t but, as we could see in the results, the choice of threshold distance and the number of initial regions can be assisted, even automated, with the analysis of the behavior of number of classes as function of threshold and number of initial regions. To do that analysis the Mahalanobis distance have showed more appropriate, as we could see in the results of figures 6 to 9.

We have made tests with others data sets with different dimensions and configurations and the results were analogs. In most cases the Mahalanobis distance help to find the threshold distance and the number of initial regions more easily than using Kullback-Leubler divergence. In some cases the Kullback-Leibler was better that Mahalanobis distance, and we conclude that in a way or another the task of choosing the threshold and the number of initial regions can be aided or automated.

References:

[1] Allan de M. Martins, Adrião D. D. Neto, and Jorge D. de Melo. Neural networks applied to classification of data based on Mahalanobis metrics. *IEEE International Joint Conference on Neural Networks*, 2003.

[2] J. B. MacQueen. Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics*, 1, 1967.
[3] Simon Haykin. *Neural Networks a Comprehensive Foundation*. Prentice Hall, second edition, 1999.
[4] T. Kohonen. *Self-Organization and Associative Memory*. Springer Verlag, 3 edition, 1989.
[5] Zhexue Huang and Michael K. Ng. A fuzzy kmeans algorithm for clustering categorical data. *IEEE Transactions on Fuzzy Systems*, 7, 1999.
[6] Jose Alfredo Costa. *Classificação Automática e Analise de Dados por Redes Neurais Auto-organizáveis*. PhD thesis, 1999. (in Portuguese)
[7] A.Ultsch and C.Vetter. Self-organizing-feature maps versus statistical clustering methods: A benchmark. *FG Neuroinformatik & Knstliche Intelligenz*, 1994.
[8] Brian S. Everitt. *Cluster Analysis*. Arnold, 1993.
[9] Allan de M. Martins, Adrião D. D. Neto, and Jorge D. de Melo. A neural network algorithm for complex pattern classification problems. *VI Brazilian Conference on Neural Networks*, 2003.
[10] H. H. Bock. Automatische classification. 1974.
[11] A. Papoulis. *Probability, Random Variables and Stochastic Processes*. Mc Graw Hill, 3 edition, 1991.
[12] Richard O. Duda , Peter E. Hart and David G. Stork, *Pattern Classification*. Wiley & Sons Inc, second edition ,2003