Unsupervised Algorithms for Vector Quantization: Evaluation on an environmental data set

Laurent Bougrain, Frédéric Alexandre

Unsupervised Algorithms for Vector Quantization: Evaluation on environmental data set

Laurent Bougrain, Frédéric Alexandre

LORIA - INRIA Lorraine, Bâtiment LORIA, Campus scientifique B.P. 239,
54506 Vandœuvre-lès-Nancy Cedex, France - e-mail: bougrain@loria.fr, falex@loria.fr

Abstract. Wave propagation laws are highly linked with environmental nature (city, country, mountains, etc...). Within the framework of a cell net planning in radiocommunication, we are interested in determining classes, homogeneous enough, upon which specific prediction models of radio electrical field can be applied. Various algorithms for unsupervised vector quantization exist and do not yield exactly the same result on the same problem because quantization can be done from different points of view. To better understand this phenomenon, this article presents evaluation of unsupervised neural networks, among the most useful for quantization, applied to a real-world problem. A particular interest is given to techniques that improve data analysis. The use of Mahalanobis' distance allows an assignment independently of the data correlation. The study of class dispersion and homogeneity using data structure and statistical analysis put in a prominent the global properties of each algorithm. Finally, we discuss the interest of these methods on a real problem of clustering linked to radiocommunication.

1 Introduction

Faced with the explosion of mobile communication systems, cell net planning is a strategic stage for telecommunication operators. Choosing location and size of the glazed zone of transmitting stations is a key point to optimize the development of a radio mobile net. Cells planning depends on the dying down of radio electrical wave. Moreover the laws of wave propagation change with the environment. There is no common theoretical model, so we have to define a partition of the environment, homogeneous enough to have a correct predictive model of the radio electrical wave.

This problem allows to compare unsupervised algorithms on real data set. We use a national geographic database which describes overground in France. We extract a random corpus of 5,000 patterns from four typical regions representing 65,000 patterns. Each pattern has 8 values: altitude and the percentage of presence in an area of four hundred side meters of seven parameters (water, wood, field, rock and 3 grades of construction density). These values are normalized according to mean and standard deviation calculated on the 65,000 patterns.

2 Models

For such tasks, unsupervised learning is very useful. This technique can cluster data without heuristic or knowledge. The wide range of methods shows that there is no algorithm available for any problem with good result. So what is their specificity ? The methods presented in this paper come from biological observation, mathematics or statistical physic. Some of them preserve topological information, or do not need to specify a number of classes, or control the probability density of selection of the vector quantization. We observe their results on our database and discuss later their properties.

2.1 Self-organization map

Self-organization maps (SOM) introduced by Kohonen are considered as the base for neural networks that use a competitive unsupervised learning. SOM expresses topological links from the inputs onto the outputs [4]. The algorithm updates prototype weights included in neighborhood of the nearest prototype of the current pattern. Neighborhood and learning rate decrease with time.

2.2 Desieno's algorithm

The competition used by SOM is called winner-take-all. It implies that some prototypes never win and are forgotten. Furthermore, prototypes in the neighborhood of the winner are updated independently of the distance to the current pattern. The prototypes stay around the center of gravity of the input patterns [6]. From theses remarks, Desieno proposes a new algorithm [2]. For each prototype, a conscience factor based on frequency of victory is added to its similarity as a handicap. It prevents the prototype selection from unequal probabilities so prototypes have the same probability. On the other hand, topological information is not preserved.

2.3 Neural Gas

Topological information of the input space is a factor we want to keep. To have it in a different way than SOM, we use Neural Gas network proposed by Martinetz [5]. Topology is not fixed as in SOM but learned. Number of connections can be parametred.

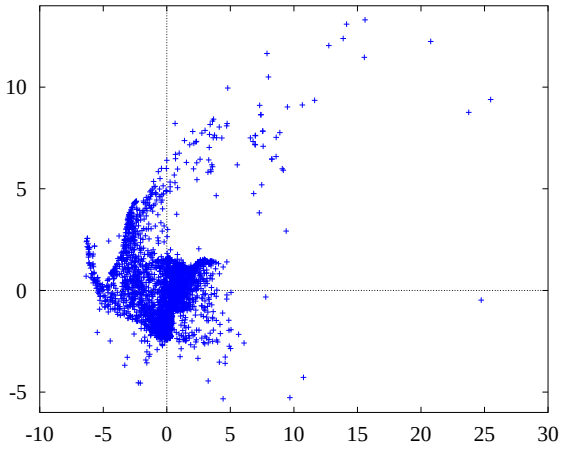

Fig. 1. Samples

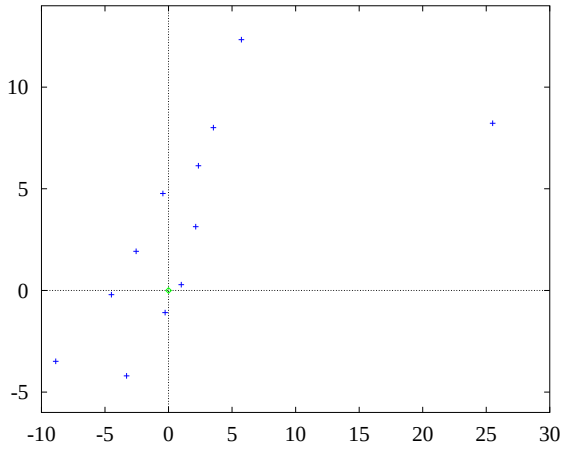

Fig. 2. Buhmann's algorithm

2.4 Growing Neural Gas

Fritzke started from Neural Gas network and does not fix the number of classes but conserves an evolutive topology [3]. Connections and prototypes are not updated exactly like Neural Gas. The number of prototypes increases up to the maximum number of classes. Stopping criterion can be different. In this case, number of classes is not fixed.

2.5 Buhmann's algorithm

Based on the algorithm of simulated annealing, which avoids local minimums, Buhmann and Kühnel proposed a vector quantization algorithm that gave an optimal number of classes depending on two parameters [1]. A new prototype is created if a high number of patterns are assigned badly or if few patterns are far away from every prototypes.

To better understand the functioning and the specificities of these unsupervised algorithms, we applied all of them on the 5,000 patterns mentionned above. Their evaluation was permitted with 2 statistical tools that we present now.

3 Evaluation

3.1 Mahalanobis' distance

The aim of classification methods is to build a partition of a set of elements gathered by proximity to get classes as homogeneous as possible. When the best k-class sharing out representation of n-elements is found, we try to assign a new element to a class. A classical rule is to compute distances between the element to be classified and reference elements. So we need to define a distance measure.

Mahalanobis' distance considers parameter correlation [8] (specific parameter dispersion is normalized by centering and data reduction). This distance between two elements $\mathbf{x}_i$ et $\mathbf{x}_j$ is defined by $d^2(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i - \mathbf{x}_j)^T \mathbf{\Sigma}^{-1} (\mathbf{x}_i - \mathbf{x}_j)$ with $\mathbf{\Sigma}$ the p-dimension square matrix equal to variance-covariance matrix of elements.
Two uses of Mahalanobis' distance have been applied. First, correlation between parameters is calculated for data set. Second, correlation are calculated within each class. There is a normalized variance-covariance matrix for each class $\mathbf{\Sigma}_k^- 1$. In practice, neural networks trained with a global variance-covariance matrix gave better results because their generalization capacity is higher.

3.2 Sammon's non linear mapping algorithm

It is difficult to estimate what the result is in a space dimension larger than 3. A principal component analysis would reduce the number of dimensions to 2 or 3 but the projection will be made without consideration for eigenvalue ie for lost information quantity (concerning our 8-dimension database, in a 3-dimension space 40% of information is lost and in a 2-dimension space 54%) . Furthermore, it is a linear projection. Sammon suggests a non linear projection to answer this problem [7]. Its algorithm tries to determine data structure to keep inter-point distance ratio after projection. All patterns from learning corpus and all prototypes from every networks are projected together in a 2-dimension space. Before projection, identical points are deleted to prevent a blow up combination.
On our database, Fig.1 shows the visual window is determined by pattern location zone. This window is kept to visualize prototypes in the aim to avoid specific zoom for each prototype set and then compare more easily the spread of prototypes from one method to other one. On these data,

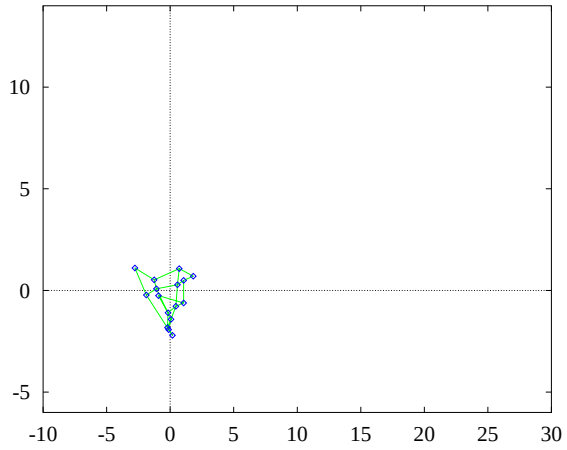

Fig. 3. SOM

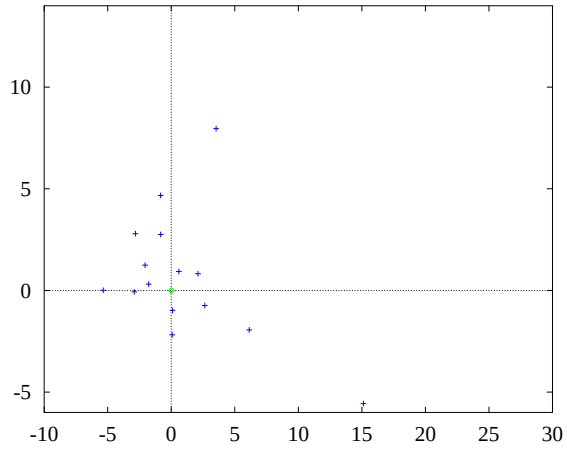

Fig. 4. Desieno's algorithm

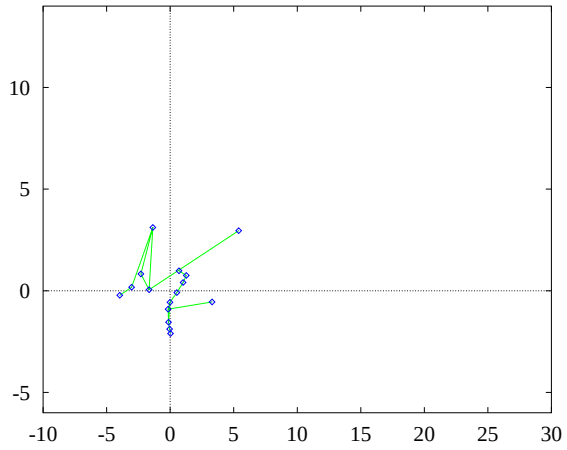

Fig. 5. Neural gas

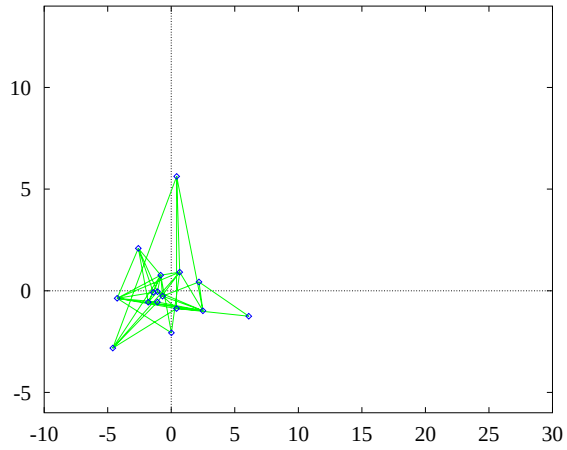

Fig. 6. Growing Neural Gas

Buhmann's algorithm create a very good spread up of prototypes in all the input space (Fig 2). Parameters were adjusted to obtain 16 classes and to compare with others neural networks. Parameters variation shows that the system stops with about 10 classes or more than 50 classes. Each SOM's prototype stays in the center of the higher density zones (Fig 3) because the winner takes its neighbors with it when it is updated. Equal probability of prototypes generated by Desieno's algorithm is a compromise solution between both of the previous methods (Fig 4). Neural gas and Growing Neural Gas spread up their prototypes in the high density zone (Fig 5 et Fig 6).

3.3 Statistical evaluation

Sammon's projection allows to visualize prototype position in comparison with patterns to better understand specifi city and strategy of each method. Nevertheless, it does not consider a pattern assignment in a class because the assignment probability in a class intervenes independently of the distance between pattern and prototype (this is the case within Buhmann's algorithm). For real values, high density zones do not appear necessarily because they are not random and certainly distinct. In the application, particular values common to several patterns represent 22%.

Statistical clustering methods can be used to estimate if the pattern assignment is homogeneous within classes, a criterion based upon intraclass inertia minimization $I = \frac{1}{2}\sum_k \sum_{i,j \in C_k} P_i P_j d^2(x_i, x_j)$ (or interclass inertia maximization).

On our application, we computed for each model: kohonen(106), Desieno(65), NeuralGas(87), GrowingNeuralGas(92), Buhmann(31). This latter result can be explained

because intraclass inertia minimization is one of the criterion used by Buhmann's algorithm to optimize its solution.

4 Conclusion

Our work evaluates some of the most useful unsupervised neural networks on a partition problem. All vector quantization methods we have tested use a distance measure between patterns. We applied Mahalanobis' distance to decorrelate parameters when we assign pattern to a class by calculating variance-covariance matrix for all patterns. To have a visual information, we projected onto a 2-dimension space, patterns and prototypes, with Sammon's non linear mapping algorithm. We can say that self-organisation maps of Kohonen is a quantitative approach. It gathered its prototypes in the center of high density zones. Weights must be initialized closely to the input patterns to prevent very different prototype probabilities. The algorithm has good result against noise because far away patterns from mean values do not attract prototypes more than the others. Desieno's algorithm is a qualitative approach. All prototypes have the same probability of assignment. Finally, all patterns have the same interest but there is no information about topology. Neural Gas et Growing Neural Gas networks have quite the same behavior. They give a compromise solution between both of the previous methods. Topology is learned and does not force the updating. Their prototypes are spread in high density zones. Buhmann's prototypes are spread out on all input space. The number of classes is not previously determined. It is optimized for an approximated quantity. On the other hand, dispersion within classes indicated that Buhmann's algorithm is a good criterion to minimize it. Methods with very unequal assignment probabilities allow to abandon classes with too small assignment probability if this is needed. The diversity of unsupervised neural networks echoes the diversity of problems. We think that such an analysis could be useful to determine which algorithm we could choose for a given problem.

5 Acknowledgement

This research was supported by CNET through Contract n^o 97 1B008.

References

1. J. Buhmann and H. Kuhnel. Complexity optimized vector quantization: A neural network approach. *Proceeding of data compression conference*, 1992.
2. D. Desieno. Adding a conscience to competitive learning. *IEEE International Conference on Neural Networks*, 1:117–124, 1989.
3. B. Fritzke. A growing neural gas network learns topologies. In G. Tesauro, D. S. Touretzky, and T. K. Leen, editors, *Advances in Neural Information Processing Systems 7*, pages 625–632. MIT Press, Cambridge MA, 1995.
4. T. Kohonen. *Self-Organization and Associative Memory*. Springer-Verlag, Berlin, 3 edition, 1989.
5. T. M. Martinetz. Competitive hebbian learning rule froms perfectly topology preserving maps. In *ICANN'93: International Conference on Artificial Neural Networks*, pages 427–434, Amsterdam, 1993. Springer.
6. D.E. Rumelhart and D. Zipser. Feature discovery by competitive learning. *Cognitive Science*, 9:75–112, 1985.
7. John W. Jr Sammon. A nonlinear mapping for data structure analysis. *IEEE Transactions on Computers*, 5:401–409, 1969.
8. G. Saporta. *Analyse des données et statistique*. Ed. Technip, 1990.