A comparative study of local classifiers based on clustering techniques and one-layer neural networks *

Yuridia Gago-Pallares, Oscar Fontenla-Romero and Amparo Alonso-Betanzos

University of A Coruña, Department of Computer Science, 15071 A Coruña, Spain
yuridiagago@hotmail.com, ofontenla@udc.es, ciamparo@udc.es

Abstract. In this article different approximations of a local classifier algorithm are described and compared. The classification algorithm is composed by two different steps. The first one consists on the clustering of the input data by means of three different techniques, specifically a k-means algorithm, a Growing Neural Gas (GNG) and a Self-Organizing Map (SOM). The groups of data obtained are the input to the second step of the classifier, that is composed of a set of one-layer neural networks which aim is to fit a local model for each cluster. The three different approaches used in the first step are compared regarding several parameters such as its dependence on the initial state, the number of nodes employed and its performance. In order to carry out the comparative study, two artificial and three real benchmark data sets were employed.

1 Introduction

Both function approximation and classification problems can be approached in one of two general ways: (a) constructing a global model that tries to fit all the input data points, or (b) dividing the input data points into several groups and learning a separate model that tries to fit in each of the local patches [3]. There are several well-known algorithms for local modelling, such as Adaptive Resonance Theory (ART) [4], Self-Organizing Maps (SOM) [5], or Radial Basis Functions (RBF) [6]. One of the first proposals of this kind of methods was the Counterpropagation model [1], that is a three-layer feed-forward network based on a Kohonen linear associator and Grossberg outstar neurons. Other representative work is the K-Winner Machine (KWM) which selects among a family of classifiers the specific configuration that minimizes the expected generalization error [2]. In training, KWM uses unsupervised Vector Quantization and subsequent calibration to label data-space partitions. In those cases in which the input data is clearly non evenly distributed, local modelling can significantly improve the overall performance. So, the best approach in these cases will consist of training a learning system for each subset of patterns that could be detected in the

* This work has been funded in part by projects PGIDT05TIC10502PR of the Xunta de Galicia and TIN2006-02402 of the Ministerio de Educación y Ciencia, Spain (partially supported by the European Union ERDF).

input data. However, one of the problems with local models is their recognition speed [7].

On the other hand, there are many learning methods for neural networks. One of the most popular is the backpropagation algorithm [8] for feedforward neural networks. This method however, has its drawbacks, namely possible convergence to local minima and slow learning speed. Several algorithms have been proposed in order to mitigate or eliminate these limitations [9, 10]. One of this proposals for one-layer neural networks with non-linear activation functions [11] is based on linear least-squares and minimizes the mean squared error (MSE) before the output nonlinearity and a modified desired output, which is exactly the actual desired output passed through the inverse of the nonlinearity. This solution leads to the obtaining of a global optimum by solving linear systems of equations, and thus using much less computational power than with standard methods. This possibility of rapid convergence to global minimum can be taken as an important advantage for the construction of a local model once the input data is already clustered by a previous algorithm.

In this paper, a local model for classification that is composed by a clustering algorithm and a subsequent one-layer neural network of the type described above is presented. Three different clustering algorithms were tried, a k-means algorithm, a Growing Neural Gas (GNG) and a Self-Organizing Map (SOM). Their results are compared in several aspects, such as as dependence on the initial state, number of nodes employed and performance. For the comparison studies, five datasets were used. Two of them were artificially generated datasets and the other three correspond to real benchmark datasets from the UCI Learning Repository [12].

2 The Local Model

The local model designed for classification problems is composed by two different algorithms. First, a clustering method is in charge of dividing the input data set into several groups, and subsequently each of these groups is fed to a one-layer neural network of the type described in [11] with logistic transfer functions, that constructs a local model for each of them. Beside the advantages of obtaining the global minimum and a rapid speed of converge, already mentioned in the Introduction section, these neural networks conform an incremental learning. This characteristic can be of vital importance in distributed environments, such as for example in learning new signature attacks in a misuse intrusion detection system. The clustering method was implemented with three different approaches, k-means, GNG and SOM, which will be briefly described in section 2.1. Also, the construction of the local learning model has two different phases: a training phase and a testing phase, which will be described in section 2.3.

2.1 The clustering algorithms implemented

As mentioned above, three different clustering algorithms were tried: k-means, GNG and SOM.

The k-means algorithm: The k-means algorithm [13] is a well-known supervised algorithm that divides the input data in k classes. It is simple to implement and works reasonably well. The main disadvantage is its dependence with the initial state, and the adjusting parameter is the number of nodes of the algorithm.

The Growing Neural Gas (GNG): The GNG is a self-organizing neural network initially proposed by Fritzke [14]. It is a very flexible structure that allows for adding/eliminating nodes of the network in execution time. The nodes of the algorithm cover all data of the input dataset, and so a modification of the original algorithm is proposed in order to adapt it for our classification purposes. The aim of the modified algorithm, published in [15], is that the nodes of the network situate in the decision region, that is, behave as the decision boundary between the classes. In this way, the modified GNG algorithm creates subregions that later could be treated by local linear classifiers. The parameters to be adjusted for this algorithm are the number of nodes, ew (ratio to adapt the nodes of the map), and lambda (node insertion frequency in the network).

The Self-Organizing Map (SOM): The Self-Organizing Map (SOM) [5] is a non-supervised and competitive learning model. The SOM defines an ordered mapping, that is a projection from a set of given input data onto a regular and usually two-dimensional grid. A data item will be mapped into the node of the map whose model is most similar to the data item, that is, has the smallest distance from the data item using a given metric, in our case, a euclidean distance. In this case, the parameters to be adjusted are the number of nodes, the number of training steps (trainlen) and four constants (a,b,c,d) that determine the vector of learning ratios and the values for the learning ratios.

2.2 The one-layer neural networks

The second phase of the system is composed by a set of one-layer feedforward neural networks, one for each cluster obtained in the previous step. The goal of each network is to fit locally the data points associated to each cluster. These networks were trained using a new supervised learning method proposed in [11]. The novelty of this approach is based on the use of an alternative cost function, similar to the classical mean-squared error (MSE) function, but measuring the errors before the nonlinear activation functions and scaling them according to the corresponding operation point of the nonlinearity. The advantage of this alternative cost function is that the global optimum can be easily obtained deriving it with respect to the weights of the network and equaling these derivatives to zero. In that case, a system of linear equations can be resolved to obtain the optimal weights of the neural network. The time consumed by this method obtaining the solution is considerably lesser than that needed by the fast iterative learning methods usually employed for neural networks (quasi-Newton, conjugate gradient, etc.). This is due to the fact that method is not an iterative algorithm but rather acquires the solution in just one step.

2.3 The construction of the local model

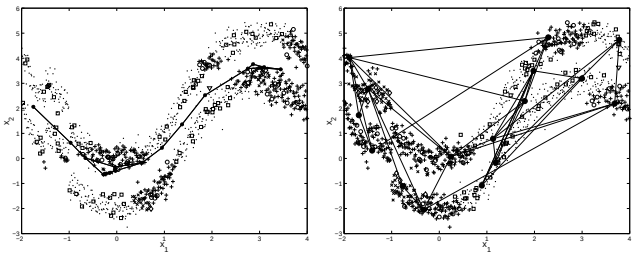
In order to obtain a final model, two different phases are needed: a training phase and a testing phase. Thus, all the datasets employed in this study were divided in training and testing sets. The training phase consists, in turn, of two stages. In the first stage, the training dataset is divided in several clusters using one of the clustering methods described in section 2.1. After this stage, several groups characterized by its centers are obtained. The second stage consists in training a set of one-layer neural networks, each one associated with one of the clusters of the previous stage. The input data for each network is the associated cluster of the first stage. The aim of this second stage is to obtain the best models for each of the clusters formed in the first stage. The testing phase consists also of two stages. In the first one, each new input data of the testing dataset is classified using the clustering method, and then, subsequently, in a second stage, that input is processed by the corresponding one-layer neural network that is determined for its group in the previous training phase.

3 Results

In order to carry out the comparative study, the three clustering algorithms were tried over five different data sets: two were artificially generated and three are benchmark data sets. The results on the tables below are obtained using a training (80% of the data set) and a test set (20% of the data set) and a further 10-fold cross-validation scheme over the training set (training set and validation set) to select the best model.

3.1 Artificial data sets

Data set 1: This data set contains two classes with a sinusoidal distribution plus some random noise. The set is composed by 1204 patterns, homogeneously distributed between the two classes. The input attributes correspond to the spatial coordinates of the points. Figure 1 shows the distribution of the data points in the input space with each class being represented by a different symbol.



(a) (b)

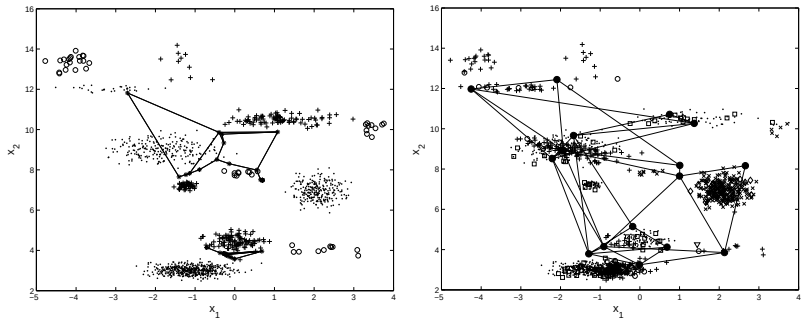
Fig. 1. Distribution of the nodes in the maps obtained for the clustering methods (a) GNG, (b) SOM, for the artificial data set 1.

Table 1 shows the percentage of error of the three implemented clustering algorithms, for the different sets, employing different configurations of the parameters. In this case all the methods obtains similar results over the test set. Morevover, figure 1 shows graphically the distribution of the nodes in the maps obtained by the modified GNG and SOM methods. As it can be observed in this figure the behaviour of both methods is completely different. While the SOM nodes try to expand over the whole dataset, the GNG nodes situate around the decision region that separates both classes.

Method	Parameters	Training (%)	Validation (%)	Test (%)
k-means	3	0.52	1.09	1.63
	8	0.04	0.20	0.29
	12	0.02	0.20	0.23
	20	0.07	0.27	0.29
	30	0.02	1.09	1.00
GNG	7/0.02/2500	1.88	2.19	3.63
	8/0.0005/5000	0.60	0.78	0.13
	20/0.0005/5000	0.22	0.47	0.50
	30/0.0005/5000	0.19	0.94	0.06
SOM	4/50/1/0.002/0.125/0.005	3.81	3.59	3.56
	9/50/1/0.002/0.125/0.005	0.22	0.44	0.50
	16/100/1/0.00004/0.000125/0.00125	0.23	0.93	0.44
	25/100/1/0.00004/0.000125/0.00125	0.12	0.16	0.69
	36/50/1/0.002/0.125/0.005	0.03	0.31	0.19

Table 1. Percentage of error of each method for the artificial data 1. The parameter for k-means is the number of nodes; for GNG Number of nodes/ew/lambda; and for SOM Number of nodes/trainlen/a/b/c/d.

Data set 2: The second artificial data set is illustrated in figure 2. In this case, there are three classes and the distribution of the data set in the input space is not homogeneous. Moreover, the number of data points in each class in not balanced (class 1: 1000 patterns, class 2: 250 patterns and class 3: 50 patterns). Table 2 shows the percentage of error of the three implemented clustering algorithms. Again, all the methods present a similar performance over the test set. Furthermore, figure 2 shows graphically the distribution of the nodes in the maps obtained by the GNG and SOM methods. As in the previous data set, the modified GNG situates the nodes of the map around the decision region. Moreover, the map is divided in two different regions: one in the center of the figure and the other on the bottom.



(a) (b)

Fig. 2. Distribution of the nodes in the maps obtained for the clustering methods (a) GNG, (b) SOM, for the artificial data set 2.

Method	Parameters	Training (%)	Validation (%)	Test (%)
k-means	5	0.84	0.67	1.2
	9	0.46	0.75	0.77
	15	0.45	0.75	0.80
	25	0.55	1.00	1.60
	36	0.56	1.25	1.77
GNG	5/0.0005/5000	0.40	1.08	0.53
	9/0.0005/5000	0.51	0.67	0.40
	15/0.0005/5000	0.23	0.50	0.23
	25/0.0005/5000	0.21	0.33	1.03
	36/0.0005/5000	0.18	0.67	0.80
SOM	4/100/0.00002/0.0000625/0.000625	0.84	0.92	0.63
	9/100/0.00002/0.0000625/0.000625	0.29	0.67	0.30
	16/100/0.00002/0.0000625/0.000625	0.17	0.42	0.30
	25/100/0.00002/0.0000625/0.000625	0.11	0.17	0.17
	36/100/0.00002/0.0000625/0.000625	0.09	0.25	0.23

Table 2. Percentage of error of each method for the artificial data 2. The parameter for k-means is the number of nodes; for GNG Number of nodes/ew/lambda; and for SOM Number of nodes/trainlen/a/b/c/d.

3.2 Real data sets

Pima Diabetes Dataset: The first real data set is the well-known Pima Indian Diabetes database. This data set contains 768 instances and 8 attributes for each class. In this case, it is not possible to represent graphically the distribution of the data and the nodes because it is not a small dimensional space. Table 3 shows the percentage of error of the three clustering algorithms for the different sets. In this case, the best results are obtained by the k-means algorithm although

the differences are not significant. Moreover, all the methods present a good behaviour in terms of generalization ability. Evidently, if many nodes or clusters are used the methods tend to overfit the training data and the generalization is affected.

Method	Parameters	Training (%)	Validation (%)	Test (%)
k-means	4	21.23	22.62	24.42
	8	21.21	21.82	26.36
	9	22.42	24.58	21.62
	20	21.77	24.60	21.56
GNG	4/0.00005/100	20.47	22.14	27.79
	8/0.0005/2500	20.21	26.55	20.97
	20/0.0005/2500	18.99	27.85	25.19
SOM	4/100/1/0.00004/0.000125/0.00125	20.83	24.08	22.40
	9/100/1/0.00004/0.000125/0.00125	18.02	24.57	30.65
	16/100/1/0.00004/0.000125/0.00125	16.81	27.38	27.59
	25/100/1/0.00004/0.000125/0.00125	14.37	27.51	31.17
	36/100/1/0.00004/0.000125/0.00125	12.03	29.13	29.74

Table 3. Percentage of error of each method for the Pima Indian Diabetes dataset. The parameter for k-means is the number of nodes; for GNG Number of nodes/ew/lambda; and for SOM Number of nodes/trainlen/a/b/c/d.

Wisconsin Breast Cancer Dataset: This database was obtained from the University of Wisconsin Hospitals, Madison from Dr. William H. Wolberg. It contains 699 instances, 458 benign and 241 malignant cases, but 16 were not used as they contain incomplete information. Each example is characterized by 9 attributes measured in a discrete range between 1 and 10. Table 4 shows the percentage of error of the three analyzed algorithms. For this database, the results over the test set are very similar for the best model of each type (selected using the validation set).

StatLog Dataset: This last dataset contains images acquired by the Landsat satellite in different frequency bands (two from the visible spectra and other two from the infrared spectra). This database has 4435 patterns for the training process and 2000 for testing. Each data point is characterized by 36 attributes (the four spectras $\times$ 9 pixels from the neighbourhood). The patterns must be categorized in 6 classes. Table 5 shows the percentage of error for each method. In this data set the k-means presents a performance significantly worse than the other methods. Again as in the previous cases, the analyzed methods present a good generalization ability.

4 Comparative results with other methods

Finally, in this section a comparative study with other machine learning methods is included. The results were obtained from [16]. These previous results were published using a 10-fold cross validation over the whole data set, thus the best

Method	Parameters	Training (%)	Validation (%)	Test (%)
k-means	2	3.54	4.38	5.18
	5	3.54	4.62	4.26
	8	3.59	4.60	3.95
	20	4.25	5.05	6.11
GNG	2/0.0005/2500	2.80	4.41	4.30
	5/0.0005/2500	1.42	5.28	7.11
	8/0.0005/2500	1.66	7.67	5.26
	20/0.0005/2500	0.95	5.92	9.12
SOM	4/100/1/0.00004/0.000125/0.00125	2.54	4.83	4.39
	9/100/1/0.00004/0.000125/0.00125	1.22	6.14	7.37
	16/100/1/0.00004/0.000125/0.00125	0.65	9.00	6.40
	25/100/1/0.00004/0.000125/0.00125	0.07	8.56	9.47
	36/100/1/0.00004/0.000125/0.00125	0.02	10.76	11.05

Table 4. Percentage of error of each method for Breast Cancer dataset. The parameter for k-means is the number of nodes; for GNG Number of nodes/ew/lambda; and for SOM Number of nodes/trainlen/a/b/c/d.

Method	Parameters	Training (%)	Validation (%)	Test (%)
k-means	3	22.00	22.75	22.02
	8	21.73	22.27	23.57
	15	23.17	23.59	22.32
	20	22.35	22.94	25.01
	38	23.01	23.56	23.71
GNG	3/0.0002/1000	12.15	13.75	14.27
	7/0.00003/900	11.24	14.43	13.44
	8/0.0005/2500	10.56	13.47	14.63
	38/0.0005/1000	6.69	15.22	15.03
SOM	4/100/1/0.00004/0.000125/0.00125	13.30	15.50	13.29
	9/100/1/0.00004/0.000125/0.00125	10.49	13.30	13.03
	16/100/1/0.00004/0.000125/0.00125	8.15	13.53	12.46
	25/100/1/0.00004/0.000125/0.00125	6.10	13.58	12.74
	36/100/1/0.00004/0.000125/0.00125	4.29	12.96	15.30

Table 5. Percentage of error of each method for the Statlog dataset. The parameter for k-means is the number of nodes; for GNG Number of nodes/ew/lambda; and for SOM Number of nodes/trainlen/a/b/c/d.

model, selected in the previous section for each data set, was retrained using this scheme in order to carried out a comparable analysis. Table 6 shows the accuracy (over the test set) of several methods, included the best results obtained by the approaches described in this paper, for the Pima Diabetes, Wisconsin Breast Cancer and Statlog datasets, respectively. As can be seen, the described methods obtain comparable results and in the case of the Pima Diabetes database they are among the best classifiers.

Dataset	Method	Accuracy
Pima	**k-means+1NN**	**77.87**
	Linear discriminant Analysis (LDA)	77.5-77.2
	GNG+1NN	**77.33**
	MLP+Backpropagtion	76.40
	SOM+1NN	**76.05**
	Learning vector quantization (LVQ)	75.80
	RBF	75.70
	C4.5	73.00
	Kohonen	72.70
WdbcData	**k-means+1NN**	**77.87**
	Linear discriminant Analysis (LDA)	77.5-77.2
	GNG+1NN	**77.33**
	MLP+Backpropagtion	76.40
	SOM+1NN	**76.05**
	Learning vector quantization (LVQ)	75.80
	RBF	75.70
	C4.5	73.00
	Kohonen	72.70
Statlog	Support vector machine (SVM)	97.20
	Fisher linear discriminant analysis	96.80
	MLP+Backpropagation	96.70
	Learning vector quantization (LVQ)	96.60
	k-means+1NN	**96.31**
	GNG+1NN	**96.31**
	Linear discriminant analysis (LDA)	96.00
	SOM+1NN	**95.96**
	RBF	95.90
	C4.5	93.40

Table 6. Comparative performance for PimaData, WdbcData and StatlogData

5 Conclusions

A two step local model classifier has been described in this work. The first step of the classifier is a clustering algorithm, and the second is a one-layer neural network which fits a local model for each of the clusters obtained by the first step. The one-layer neural network is of a type that obtains a global optimum by solving linear systems of equations, and thus using much less computational power than with standard methods. Three different algorithms were used for

the clustering phase: k-means algorithm, GNG and SOM. The local models were tested over five different datasets. The performance obtained by the models were adequate, and in fact, are among the best results obtained when compared with other machine learning methods, as it can be seen on the previous section. In fact, for the case of the Pima Indian Diabetes dataset, our methods are the ones that obtain the best results. Regarding the three different clustering algorithm, k-means is, in average, the one that obtains worst results, probably due to its high dependence on the initial state. The other two clustering methods, GNG and SOM, are the ones obtaining best results, although their performance depends on several parameters is high, and so they are difficult to adjust. Finally, the generalization ability of all the analyzed methods is good, however it can be affected if many local models are used in the clustering phase.

References

1. Hecht-Nielsen, R.: Neurocomputing. Addison-Wesley, Reading, MA (1990)
2. Ridella, S., Rovetta, S., Zunino, R.: The k-winner machine model. In: Int. Joint Conference on Neural Networks (IJCNN00). Volume 1., IEEE (2000) 106–111
3. Alpaydin, E.: Introduction to Machine Learning. MIT Press (2004)
4. Carpenter, G., Grossberg, S.: The art of adaptive pattern recognition by a self-organizing neural network. IEEE Computer **21**(3) (1988) 77–88
5. Kohonen, T.: Self-organizing maps. Springer (2001)
6. Broomhead, D., Lowe, D.: Multivariable functional interpolation and adaptive networks. Computer Systems **2** (1988) 321–355
7. Bottou, L., Vapnik, V.: Local learning algorithms. Neural Computation **4** (1992) 888–900
8. Rumelhart, D.E., Hinton, G.E., Willian, R.J.: Learning representations of back-propagation errors. Nature **323** (1986) 533–536
9. Moller, M.F.: A scaled conjugate gradient algorithm for fast supervised learning. Neural Networks **6** (1993) 525–533
10. Hagan, M.T., Menhaj, M.: Training feedforward networks with the Marquardt algorithm. IEEE Transactions on Neural Networks **5**(6) (1994) 989–993
11. Castillo, E., Fontenla-Romero, O., Alonso-Betanzos, A., Guijarro-Berdiñas, B.: A global optimum approach for one-layer neural networks. Neural Computation **14**(6) (2002) 1429–1449
12. Newman, D., Hettich, S., Blake, C., Merz, C.: UCI repository of machine learning databases. http://www.ics.uci.edu/∼mlearn/MLRepository.html (1998)
13. MacQueen, J.B.: Some methods for classification and analysis of multivariate observations. In: 5-th Berkeley Symposium on Mathematical Statistics and Probability. Volume 1., Berkeley, University of California Press (1967) 281–297
14. Fritzke, B.: A growing neural gas network learns topologies. In Tesauro, G., Touretzky, D.S., Leen, T.K., eds.: Advances in Neural Information Processing Systems 7 (NIPS'94), Cambridge, MA, MIT Press (1995) 625–632
15. Rodríguez-Pena, R.M., Pérez-Sánchez, B., Fontenla-Romero, O.: A novel local classification method using growing neural gas and proximal support vector machines. In: Int. Joint Conference on Neural Networks (IJCNN07), IEEE (2007)
16. Computational Intelligence Laboratory, Department of Informatics, Nicolaus Copernicus University: Datasets used for classification: comparison of results. (http://www.fizyka.umk.pl/kmk/projects/datasets.html (Last access: July 17, 2007))