Robust Neural Network Classifier

[1]Mohamed M. Zahra,[2]Mohamed H. Essai,[3]Ali R. Abd Ellah

Electrical and Electronics Engineering, Al-Azhar University, Qena, Egypt
[1]mzahra15455@gmail.com, [2]mhessai@azhar.edu.eg, [3]ali_refaee2000@yahoo.com

***Abstract* - Classification is a data mining technique used to predict Patterns' membership. Pattern classification involves building a function that maps the input feature space to an output space of two or more than two classes. Neural Networks (NN) are an effective tool in the field of pattern classification. The success of NN is highly dependent on the performance of the training process and hence the training algorithm. Many training algorithms have been proposed so far to improve the performance of neural networks. Usually a traditional backpropagation learning algorithm (BPLA), which minimizes the mean squared error (MSE – cost function) of the training data, be used in the process of training neural networks. However (MSE) based learning algorithm is not robust in presence of outliers that may pollute the training data. In our work we aim to present another cost functions which backpropagation learning algorithm based on in order to improve the robustness of neural network training by employing a family of robust statistics estimators, commonly known as M-estimators, and hence obtain robust NN classifiers. Comparative study between robust classifiers and non-robust (traditional) classifiers was established in paper using crab classification problem.**

***Index Terms* --- Robust Statistics, Feed-Forward Neural Networks, M-Estimators, Classification, Robust classifier.**

I. INTRODUCTION

Artificial Neural Networks have been successfully used in a number of applications due to their highly advantageous properties like parallel processing of information, capacity to handle non-linearity, quick adaptability to system dynamics, and many more. They can be trained to efficiently recognize patterns of information in the presence of noise and non-linearity, and classify the information using those patterns. These properties can be exploited to use artificial neural networks in the actively researched field of classification [1].

Advantages of neural networks, however, include their high tolerance of noisy dataas well as their ability to classify patterns on which they have not been trained. They can be used when you may have little knowledge of the relationships between attributes and classes. They are well-suited for continuous-valued inputs and outputs, unlike most decision tree algorithms, as shown…….[2]. They have been successful on a wide array of real-world data, including handwritten character recognition, pathology and laboratory medicine, and training a computer to pronounce English text. Neural network algorithms are inherently parallel; parallelization techniques can be used to speed up the computation process. In addition, several techniques have recently been developed for the extraction of rules from trained neural networks. These factors contribute toward the usefulness of neural networks for classification and prediction in data mining [2].

Feed-forward neural networks are commonly trained by the traditional back propagation learning algorithm. It is common to use the back propagation learning algorithm based on the minimization of the mean square error (MSE) for the training data. The use of (MSE) in data modeling is commonly known as the least mean squares (LMS) method. The basic idea of (LMS) is to optimize the fit of a model with respect to the training data by minimizing the square of residuals. Mean squared error (MSE) is the preferred measure in many data modeling techniques. Tradition and ease of computation account for the popularity of (MSE).

M-estimators are a class of estimators belong to robust statistics [3] family generated from the maximum likelihood estimators that are designed to be stable under minor noise perturbation and robust against gross errors in the data. We shall use it in order to robustify the NN learning process [4], [5], [6] in the presence of contaminated data (outliers). They try to reduce the effect of outliers by replacing the squared residual by another function of residual.

Outliers are sample values that are dramatically different from patterns in the rest of the data, and cause surprise in relation to the majority of samples. Outliers are a common feature in many real data sets, their occurrence in raw data ranges from 1 to 10%. They may be due to measurement error, or they may represent significant features in the data. Identifying outliers, and deciding what to do with them, depend on an understanding of the data and its source.

Classification is one of the most frequently encountered decision making tasks of human activity. A classification problem occurs when an object needs to be assigned into a predefined group or class based on a number of observed attributes related to that object. Many problems in business, science, industry, and medicine can be treated as classification problems. Examples include bankruptcy prediction, credit scoring, medical diagnosis, quality control, handwritten character recognition, and speech recognition [7].

Classification analysis is the organization of data in given classes. Also known as supervised classification, the classification uses given class labels to order the objects in the data collection. Classification approaches normally use a training set where all objects are already associated with known class labels. The classification algorithm learns from the training set and builds a model. The model is used to classify new objects [8].

The objective of our contribution in this paper is to introduce robust neural network classifiers, exploiting ANN capabilities in the field of classification, that will be trained using the robust back propagation learning algorithm, which essentially depends on M-estimators as performance functions (cost functions) instead of (MSE) performance function in order to optimize the robustness of the neural networks training, and hence NN classifier in case of corrupted (noisy)data (outliers).

The outline of this paper is as follows. Section (2) presents M-estimators and shows some common M-estimators. Section (3) back propagation learning algorithm based M-estimators. Section (4) discusses Outliers Analysis and Noisy data. Section (5) discusses the classification. Section (6) gives our experimental results by comparing the performance of various M-estimators and MSE in terms of accuracy in case of corrupted data (Outliers).

II. M-ESTIMATORS

M-estimators have gained popularity in the neural networks community [9]. Let $\mathrm{r_i}$ be the residual of the i^{th} datum, i.e. the difference between the i^{th} observation and its fitted value. The standard least-squares method tries to optimize the training data by minimize $\sum_i r_i^2$ but The M-estimators try to minimize the error by replacing the squared residuals r_i^2 by another function of the residuals, yielding

$$Min \sum_i \rho(r_i) \tag{1}$$

Where is $\rho(.)$ a symmetric, positive-definite function with a unique minimum at zero, and is chosen to be less increasing than square. Table 1, [9] lists a few commonly used M-estimators and their influence functions. M-estimators'

Table.1, Some Commonly used M-estimators

Type	$\rho(r)$	$\psi(r)$	$\omega(r)$
L2	$r^2/2$	r	1
L1	$\lvert r\rvert$	Sgn(r)	$\frac{1}{\lvert r\rvert}$
Fair	$c^2[\frac{\lvert r\rvert}{c}-\log(1+\frac{\lvert r\rvert}{c})]$	$\frac{r}{1+\lvert r\rvert/c}$	$\frac{1}{1+\lvert r\rvert/c}$
Huber $\begin{cases} if\lvert r\rvert\leq k \\ if\lvert r\geq k\rvert \end{cases}$	$\begin{cases} r^2/2 \\ k(\lvert r\rvert-k/2) \end{cases}$	$\begin{cases} r \\ k.\mathrm{sgn}(r) \end{cases}$	$\begin{cases} 1 \\ k/\lvert r\rvert \end{cases}$
Cauchy	$\frac{c^2}{2}\log(1+(r/c)^2)$	$\frac{r}{1+(r/c)^2}$	$\frac{1}{1+(r/c)^2}$
Geman-McClure	$\frac{r^2/2}{1+r^2}$	$\frac{r}{(1+r^2)^2}$	$\frac{1}{(1+r^2)^2}$
Least Mean Log Of Squares	$\log(1+\frac{1}{2}r^2)$	$\frac{r}{1+\frac{1}{2}r^2}$	$\frac{1}{1+\frac{1}{2}r^2}$

III. BACKPROPAGATION LEARNING ALGORITHM BASED M-ESTIMATORS

To implement the tradition learning algorithm based on M-estimators concept, all want to do is replacing the squared residuals r_i^2 by another function of the residuals, yielding

$$E = \sum_i \rho(r_i) \tag{2}$$

. Where ρ is asymmetric, positive definite function with a unique minimum at zero, and is chosen to be less increasing than square.

IV. OUTLIERS ANALYSIS AND NOISY DATA

A data base may contain data objects that do not comply with the general behavior ormodel of the data. These data objects are outliers. Most data mining methods discard outliers as noise or exceptions. However, in some applications such as fraud detection, the rare events can be more interesting than the more regularly occurring ones. The analysis of outlier data is referred to as outlier mining.

Outliers may be detected using statistical tests that assume a distribution or probability model for the data, or using distance measures where objects that are a substantial distance from any other cluster are considered outliers. Rather than using statistical or distance measures, deviation-based methods identify outliers by examining differences in the main characteristics of objects in a group [2].

Noisy data is meaningless data. The term has often been used as a synonym for corrupt data. However, its meaning has expanded to include any data that cannot be understood and interpreted correctly by machines, such as unstructured text. Any data that has been received, stored, or changed in such a manner that it cannot be read or used by the program that originally created it can be described as noisy.

Noisy data unnecessarily increases the amount of storage space required and can also adversely affect the results of any data mining analysis. Statistical analysis can use information gleaned from historical data to weed out noisy data and facilitate data mining. Noisy data can be caused by hardware failures, programming errors and gibberish input from speech or optical character recognition (OCR) programs [10].

V. CLASSIFICATION

Classification is the process of finding a model that describes and distinguishes data classes or concepts for the purpose of being able to use the model to predict the class of objects whose class label is unknown.

Classification model can be represented in various forms such as Neural Networks and A decision trees, etc.

Classification is a multivariate technique concerned with data cases (i.e. observations) assigning [11], [12] to one of a fixed number of possible classes (represented by nominal output variables). The goal of classification is to sort observations into two or more labeled classes. The emphasis is on deriving a rule that can be used to optimally assign new objects to the labeled classes. In statistics, where classification is often done with logistic regression or a similar procedure, the properties of observations are termed explanatory variables.

A large number of input variables can present severe problems for pattern recognition systems. One technique to alleviate such problems is to combine input variables together to make a smaller number of new variables called features.

In the terminology of pattern recognition, classifications are known as the training set and future cases form the test set and our primary measure of success is the error or (misclassification) rate.

Classification problems can be seen as particular cases of function approximation, where for classification problems the functions which we seek to approximate are the probabilities of membership of the different classes expressed as functions of the input variables. Many of the key issues which need to be addressed in tackling pattern recognition problems are concerned to classification.

The task of the classifier component proper of a full system is to use the feature vector provided by the feature extractor to assign the object to a category. Because perfect classification performance is often impossible, a more general task is to determine the probability for each of the possible categories. The abstraction provided by the feature-vector representation of the input data enables the development of a largely domain-independent theory of classification.

The degree of difficulty of the classification problem depends on the variability in the feature values for objects in the same category relative to the difference between feature values for objects in different categories.

The variability of feature values for objects in the same category may be due to complexity, and may be due to noise. We define noise in very general terms: any property of the sensed pattern, which is not due to the true underlying model but instead to randomness in the world or the sensors. All nontrivial decision and pattern recognition problems involve noise in some form.

VI. SIMULATION RESULTS

Neural networks introduced as proficient classifiers and are particularly well suited for addressing non-linear problems. Given the non-linear nature of real world problems, like crab classification, neural networks is certainly a good candidate for solving the problem.

In this section we attempt to build a classifier that can identify the sex of a crab from its physical measurements. Six physical characteristics of a crab are considered: species, frontal lip, rear width, length, width and depth [13].

For a comparison the constructed classifiers will be trained using BP learning algorithm, that each time will use one of M-estimators as a performance function in order to get robust classifier type, and the traditional (MSE) performance function in order to get traditional classifier one.

The six physical characteristics will be organized as input matrix to a neural network where i^{th}column of this matrix contains six elements representing crab’s features (species, frontal lip, rear width, length, width and depth), and the sex of the crab will be organized as target matrix, where each corresponding column of the target matrix will have two elements. Female crabs are represented with a one in the first element, and male crabs with a one in the second element. Given an input, matrix, the neural network then will be tuned to produce the desired target outputs (process of neural network training). After this process it is expected that NN will have ability to identify if the crab is male or female [13].

We shall study the performances of classifiers (robust and traditional) in three cases with respect to disturbance (outlier’s percent ε) percent of the crab's features.

Set A, Neural networks trained with high-quality data corrupted with small Gaussian noise: G_2~N (0,0.1).

Set B, Neural networks trained data corrupted with Gaussian noise, $G2$, in addition to high value random outliers of the form:
H_1~ N (-15,2), H_2~N (-20,3), H_3 ~ N (+30,1.5), H_4~ N (-12,3).

The data perturbation used in this case is as follows:

$$\text{Data} = (1-\varepsilon\%)G_2 + \varepsilon\%(H_1 + H_2 + H_3 + H_4)$$

The outliers were introduced in the data with percentage:

Set B. With $\varepsilon\%$=0.1.

Set C, Neural networks trained with 49% of the data corrupted with Gaussian noise G_2 ~ N (0,0.1); and the remaining 51% of the data substituted by background noise, uniformly distributed.

Case1: In this case we will disturb the all six crab's features, using above noisy data models (Set A, Set B, and Set C).

Table.2, Robust and traditional classifiers performance comparison
in case of disturbance of all crab's features

Performance function	percentage of correct classification t		
	Set A	Set B	Set C
MSE	96.7%	63.3%	53.3%
Geman-M clure	100%	76.7%	86.7%
Cauchy	100%	93.3%	93.3%
Lmls	100%	90%	96.7%
L1	100%	70%	80%
Fair	100%	73.3%	86.7%
Huber	63.3%	80%	70%

We studied the training of neural networks in this case as follow.

Set A - The performances of robust and traditional classifiers, represented as correct classification percentage are given in Table.2.

It is clear from Table.2, that classifiers trained using both GM, Cauchy, Lmls, L1, and Fair (robust classifiers) have identical 100% correct classification percentage, while classifier which trained using Huber has the less correct classification percentage 63.3% in comparison with its peers.

Also it is clear that the classifier (traditional classifier) which trained using MSE has a 96.7% correct classification percentage.

Set B -In the presence of corrupted data the robust classifiers which trained using Cauchy, Lmls, and Huber have the highest correct classification percentages.

The best of all is the robust classifier which trained using Cauchy, where it achieved 93.3% correct classification percentage, this percentage is not far from the percentage in Set A.

Also it is clear that the traditional classifier has the lowest correct classification percentage, which equals 63.3%.

Set C - In this case the robust classifier which trained using Lmls is the best of all with correct classification percentage 96.7%, while Cauchy, GM, Fair and L1 have percentage of correct classification is not far from Lmls.

Also it is clear that the traditional classifier has the lowest correct classification percentage, which equals 53.3%.

Case2: In this case we will disturb only three crab's features, using above noisy data models (Set A, Set B, and Set C).

Table. 3, Robust and traditional classifiers performance comparison
in case of disturbance of only three crab's features

Performance function	percentage of correct classification		
	Set A	Set B	Set C
MSE	100%	53.3%	63.3%
Geman-M clure	100%	86.7%	83.3%
Cauchy	100%	90%	83.3%
Lmls	100%	73.3%	83.3%
L1	100%	90%	86.7%
Fair	100%	73.3%	73.3%
Huber	53.3%	90%	86.7%

Set A -The performances of robust and traditional classifiers, represented as correct classification percentage are given in Table.3.

It is clear from Table.3, that classifiers trained using both GM, Cauchy, Lmls, L1, and Fair (robust classifiers) have identical 100% correct classification percentage, while classifier which trained using Huber has the less correct classification percentage 53.3% in comparison with its peers.

Also it is clear that the traditional classifier has a identical 100% correct classification percentage as robust classifiers.

Set B -In the presence of corrupted data the robust classifiers which trained using Cauchy, L1, and Huber have the highest and identical correct classification percentages of 90% percentage which is not far from the percentage in Set A.

Also it is clear that the traditional classifier has the lowest correct classification percentage, which equals 53.3%.

Set C- in this case the robust classifier trained using both L1 and Huber are the best of all with correct classification percentage 86.7%. While Cauchy, GM, Lmls, and fair have percentage of correct classification is not far from their peers.

Also it is clear that the traditional classifier has the lowest correct classification percentage, which equals 63.3%.

Case3: In this case we will disturb only one crab's feature, using above noisy data models (Set A, Set B, and Set C).

Table. 4, Robust and traditional classifiers performance comparison
in case of disturbance of only one crab’s feature.

Performance function	percentage of correct classification		
	Set A	Set B	Set C
MSE	100%	56.7%	46.7%
Geman-M clure	100%	73.3%	76.7%
Cauchy	100%	90%	86.7%
Lmls	100%	90%	86.7%
L1	100%	86.7%	73.3%
Fair	100%	80%	76.7 %
Huber	46.7%	83.3%	80%

Set A -The performances of robust and traditional classifiers, represented as correct classification percentage are given in Table.4.

It is clear from Table.4, that classifiers trained using both GM, Cauchy, Lmls, L1, and Fair (robust classifiers) have identical 100% correct classification percentage, while classifier which trained using Huber has the less correct classification percentage 46.7% in comparison with its peers.

Also it is clear that the traditional classifier has an identical 100% correct classification percentage as robust classifiers.

Set B -In the presence of corrupted data the robust classifiers which trained using Cauchy, and Lmls have the highest and identical correct classification percentages of 90% percentage. Also it is clear that the traditionalclassifier has the lowest correct classification percentage, which equals 56.7%.

Set C - In this case the robust classifier trained using both Cauchy, and Lmls are the best of all with correct classification percentage 86.7%, that trained using Huber has 80% correct classification percentage , while that trained using GM, L1, Fair, have correct classification percentages are not far from the last.

Also it is clear that the traditional classifier has the lowest correct classification percentage, which equals 46.7%.

CONCLUSION

The MSE–cost (performance) function used by most learning algorithms in training NN-classifiers works well when the data set is exact or contains only a small Gaussian noise (data Set A), in our work we were called this classifier as traditional classifier. The classifier’s predictions produced by the traditional classifier is inaccurate when the data contains outliers. The proposed robust classifiers that use M-estimators (robust statistic) technique, as performance functions on the other hand, works well whether the data set is exact, contains Gaussian noise, or has outliers.

Based on the obtained results, we strongly introduce the robust classifiers that use M-estimators as robust performance functions, to replace the traditional one, where they achieved the best correct classification percentages whether the data set contains Gaussian noise, or has outliers. Also, we strongly introduce the robust classifier that used Cauchy-M-estimator robust performance function, in the learning process, as the best of all.

REFERENCES

[1] Chintan Trivedi, Mo-Yuen Chow, Arne Nilsson and H. Joel Trussell "Classification of Internet Traffic using Artificial Neural Networks" NC State University pp 1-10, 2002

[2] Jiawei Han and Micheline Kamber "Data Mining: Concepts and Techniques" Second Edition 2006

[3] P. J. Huber. Robust Statistics. John Wiley & Sons, New York, 1981

[4] M. El-Melegy, M. Essai, and A. Ali, “Robust training of artificial feedforward neural networks,” Foundations of Computational Intelligence: Learning and Approximation, A. Hassanien, A. Abraham, A.Vasilakos, and W. Pedrycz (Eds.), Springer Studies in Computational Intelligence, vol. 1, pp. 217–242, Jun. 2009.

[5] M.T El-Melegy," RANSAC Algorithm with Sequential Probability Ratio Test for Robust Training of Feed Forward Neural Networks", IEEE, International Joint Conference on Neural Networks (IJCNN), pp 3256-3263, July 31 - August 5-2011

[6] Andrzej Rusiecki," Robust Learning Algorithm Based on Iterative Least Median of Squares" Springer, pp 145-160, 15-may-2012

[7] Guoqiang Peter Zhang "Neural Networks for Classification: A Survey" IEEE, VOL. 30, NO. 4, pp 1-12, Nov-2000

[8] Osmar R. Zaïane "Introduction to Data Mining"1999

[9] Zhengyou Zhang. Parameter Estimation Techniques : A Tutorialwith Application toConic Fitting. October-1995

[10] Chirag Patel, Atul Patel, Dharmendra Patel , " Optical Character Recognition by Open Source OCR Tool Tesseract: A Case Study" International Journal Computer Applications , Volume 55– No.10, pp 50-56, October-2012

[11] Bishop, C.M. "Neural Networks for Pattern Recognition" Oxford: Clarendon Press - 1995

[12] Ripley, B.D, "Pattern Recognition and Neural Networks. Cambridge: Cambridge University press" 1996

[13] Mark Hudson Beale, Martin T. Hagan, Howard B.Demuth, Neural networks Toolbox™ 7 User's Guide