International Journal of Computer Applications Technology and Research
Volume 2– Issue 5, 548 - 551, 2013, ISSN: 2319–8656

Data Dictionary Classification using Multilayer Artificial Neural Network with New Error Metrics

Jitendra Nath Shrivastava
Singhania University,
Jhunjhunu, Pacheri Bari,
Rajasthan, India

Maringanti Hima Bindu
Department of Computer Science and Applications,
North Orissa University, India

Abstract: The ANN technique is inspired by biological neurons. In past, Artificial Neural Networks has been used as a data classification technique. In this paper, artificial neural network is used as a data classifier. Here, new error metric are considered in the data classification. The results presented in the paper, clearly shows that ANN can acts as a very good classifier with new error metrics.

Keywords: Data Dictionary; ANN; LMS; geometric error metric, Harmonic error metric.

1. INTRODUCTION

In many applications, data classification is very necessary. In past Artificial neural network has been used as data classifier with accuracy more than 95%.[1,2] However, in some application the accuracy level of more than 99% is required. One such application is spam e-mail filtering where, data dictionary which consists of spam words need to be very accurate.

A neural network is an interconnected group of artificial neurons that uses a mathematical or computational model for information processing based on a connectionist approach [3]. Artificial neural networks are parallel computational models which are able to map any nonlinear functional relationship between an input and an output hyperspace to desired accuracy. They are constituted by individual processing units called neurons or nodes and differ among each other in the way these units are connected to process the information and, consequently in the kind of learning protocol adopted [4].

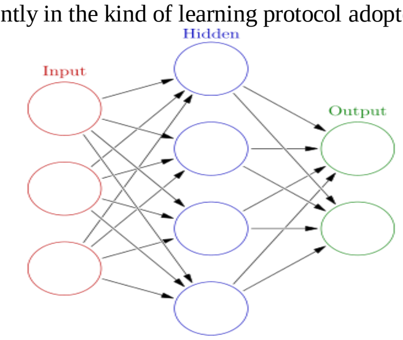

Figure 1 The Basic ANN Structure

In particular, the neurons of a feed-forward neural network are organized in three layers: the input units receive information from the outside world, usually in the form of a data file; the intermediate neurons, contained in one or more hidden layers, allow nonlinearity in the data processing, the output layer is used to provide an answer for a given set of input values [Figure 1]. In a fully connected artificial neural network, each neuron in a given layer is connected to each neuron in the following layer by an associated numerical weight (w_{ij}), the weight that passes between them. In addition, each neuron possesses a numerical bias term corresponding to an input of −1 whose associated weight has the meaning of a threshold value. Rumelhart et al. [5] popularized the use of back-propagation for learning internal representation in neural networks. Back-propagation (BP) algorithm is the most widely used search technique for training neural networks. Information in an ANN is stored in the connection weights which can be thought of as the memory of the system. The purpose of BP training is to change iteratively the weights between the neurons in a direction that minimizes the error E, defined as the squared difference between the desired and the actual outcomes of the output nodes, summed over training patterns (training set data) and the output neurons. The algorithm uses a sample-by-sample updating rule for adjusting connection weights in the network. In one algorithm iteration, a training sample is presented to the network. The signal is then fed in a forward manner through the network until the network output is obtained. The error between the actual and desired network outputs is calculated and used to adjust the connection weights.

Basically, the adjustment procedure, derived from a gradient descent method, is used to reduce the error magnitude. The procedure is firstly applied to the connection weights in the output layer, followed by the connection weights in the hidden layer next to output layer. This adjustment is continued backward through to network until connection weights in the first hidden layer are reached. The iteration is completed after all connection weights in the network have been adjusted. In this study, training of the multi-layer neural networks is implemented with back-propagation algorithm and network structure that has been trained with back-propagation algorithm has been used in the solutions of the multi-group classification models.

In Back-propagation algorithm, training of the neuron model is done by minimizing the error between target value and the observed value. In order to determine error between target and observed value, distance metric is used. It has been observed that Euclidean distance metric is the most commonly used for error measures in Neural Network applications. But it has been suggested that this distance metric is not appropriate for many problems [6]. In this work the aim is to find best error metric to use in Back propagation learning algorithm. The likelihood and log-likelihood functions are the basis for deriving estimators for parameters, for given set of data. In maximum likelihood method we estimate the value of ‘y’ for the given value of ‘x’ in presence of error. (See equation (1))

$$y_i = x_i + e \qquad (1)$$

Let x_i denote the data points in the distribution and let N denotes the number of data points. Then an estimator $\overline{\mu}$ of µ can be estimated by minimizing the error metric with respect to $\overline{\mu}$.

$$\varepsilon = \sum_{i=1}^{N} f(x, \overline{\mu}). \qquad (2)$$

Where $f(x, \overline{\mu})$ is distance metric.

Jie, et. al., [7-8] has proposed some new distance metrics based on different means. These distance metrics can be used to improve the performance of the neuron model for learning the best-fit weights of the neuron models. Distance metrics associated with the distribution models that imply the arithmetic mean, harmonic mean and geometric mean in (See Table 1) are inferred using equation

$$\frac{d\varepsilon}{d\overline{\mu}} = \frac{d}{d\overline{\mu}} \sum_{i=1}^{N} f(x, \overline{\mu}) = 0 \qquad (3)$$

In past, it is found that in the distribution associated with the harmonic and geometric estimations, the observations x_i which are far away from $\overline{\mu}$ will contribute less towards µ, in contrast to arithmetic mean and thus the estimated values will be less sensitive to the bad observations (i.e., observation with large variance), and therefore they are more robust in nature [9].

Table 1 Error Metric Types and Mean

	Error metric	**Mean**
Arithmetic	$\varepsilon = \sum_{i=1}^{N} (x_i - \mu)^2$	$\mu = \frac{1}{N} \sum_{i=1}^{N} x_i$
Harmonic	$\varepsilon = \sum_{i=1}^{N} x_i \left(\frac{\mu}{x_i} - 1 \right)^2$	$\mu = \frac{N}{\sum_{i=1}^{N} x_i}$
Geometric	$\varepsilon = \sum_{i=1}^{N} \left[\log \left(\frac{\mu}{x_i} \right) \right]^2$	$\mu = \frac{1}{N} \left(\Pi_{i=1}^{N} x_i \right)^{1/N}$

Considering three layer structure of ANN, It has n_i, n_h and n_o neurons in input, hidden and output later respectively. Le the input and output vectors are $X = [x_1, x_2, ... x_{ni}]^T$ and $Y = [y_1, y_2, ... y_{ni}]^T$ respectively. Considering that the weight of the neuron that connects the i^{th} neuron of input layer to j^{th} neuron of output layer is w_{ij} and bias of the j^{th} neuron of the hidden layer is bh_j, the net value of the j^{th} neuron is given by

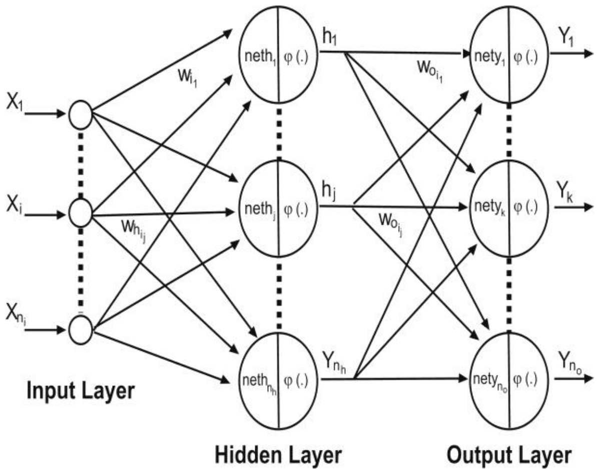

Figure 2 The Multilayer ANN Structure

$$neth_j = \sum_{i=1}^{n_i} wh_{ij}.x_i + bh_j \text{ where, } j = 1, 2 ... n_h. \qquad (4)$$

The output of the j^{th} neuron h_j of the hidden layer after applying the activation is defined as

$$h_j = \varphi(neth_j) = \frac{1}{1 + e^{-neth_j}}. \qquad (5)$$

Similarly, the net value $nety_k$ and the final output y_k of the k^{th} neuron of the output layer can be defined as

$$nety_k = \sum_{i=1}^{n_i} wo_{ij}.h_j + bh_k \text{ where, } k = 1, 2 ... n_o. \qquad (6)$$

The output of the k^{th} neuron of the output layer

$$y_k = \varphi(nety_k) = \frac{1}{1 + e^{-nety_k}} \qquad (7)$$

In this section the error back-propagation learning of MLP with different error metrics have been derived. Let E denote the cumulative error at the output layer. In BP algorithm aim is to minimize the error at the output layer. The weight update equations using gradient descent rule are given bellow:

$$wh_{ji}(n) = wh_{ji}(o) + \eta \frac{\partial E}{\partial y_k} . \frac{\partial y_k}{\partial wh_{ji}}$$

$$bh_j(n) = bh_j(o) + \eta \frac{\partial E}{\partial y_k} . \frac{\partial y_k}{\partial bh_j}$$

$$wo_{kj}(n) = wo_{kj}(o) + \eta \frac{\partial E}{\partial y_k} . \frac{\partial y_k}{\partial wo_{kj}}$$

$$bo_k(n) = bo_k(o) + \eta \frac{\partial E}{\partial y_k} . \frac{\partial y_k}{\partial bk_k} \qquad (8)$$

Where, η is learning rate and

$$\frac{\partial y_k}{\partial wh_{ji}} = \left[\sum_{k=1}^{n_0}(1-y_k).y_k.w_{kj}\right].(1-h_j).h_j.x_i$$

$$\frac{\partial y_k}{\partial bh_j} = \left[\sum_{k=1}^{n_0}(1-y_k).y_k.w_{kj}\right].(1-h_j).h_j$$

$$\frac{\partial y_k}{\partial bo_k} = \left[\sum_{k=1}^{n_0}(1-y_k).y_k\right]$$

$$\frac{\partial y_k}{\partial wo_{jk}} = \left[\sum_{k=1}^{n_0}(1-y_k).y_k.h_j\right] \qquad (9)$$

For different error criterion only $\frac{\partial E}{\partial y}$ will change and is shown as follows: This is dependent on the distance metric used in computation of the total error E.

Case 1 Least Mean Square Error

$$LMSE = E = \frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{n_o}(t_k - y_k)^2 \text{, then}$$

$$\frac{\partial E}{\partial y_k} = -\eta(t_k - y_k) \qquad (10)$$

Case 2 Geometric Error Metric

$$LMSE = E = \frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{n_o}\log\left(\frac{y_k}{t_k}\right)^2 \text{, then}$$

$$\frac{\partial E}{\partial y_k} = -\eta(\log(t_k) - \log(y_k)).\frac{1}{y_k} \qquad (11)$$

Case 3 Harmonic Error Metric

$$LMSE = E = \frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{n_o}\left(\frac{y_k}{t_k} - 1\right)^2 \text{, then}$$

$$\frac{\partial E}{\partial y_k} = -\eta.t_k.\left(\frac{y_k}{t_k} - 1\right) \qquad (12)$$

Where, y_t denotes the desired value of neuron and t_i denotes the targeted value of the i^{th} pattern.

2. RESULTS

In our experiment 421 words are considered (appendix A). These words are classified into seven groups. The group division is done in the broad categories: adult, financial, commercial, beauty and diet, travelling, home based and gambling. The ANN network is trained with 1200 words. Then the ANN used as classifier with different error metric presented in Table1. The correct classification data in percentage is shown in Table 2. It can be observed for the table that, arithmetic error type produces 96% accurate results while geometric error type produces 98% accurate results. Among the three, harmonic error metric produces 99% correct results.

Table 2: Percentage Classification under Various Error Matrixes.

Type	Correct classification in percentage
Arithmetic	96%
Geometric	98%
Harmonic	99%

3. CONCLUSIONS

In this paper, ANN based data classifier is detailed with new error metric. Here, the wordlist we call it as data dictionary 421 words are considered as we use ANN based approach to classify them into seven categories. It has been fund that harmonic error metric data classification accuracy is 99%.

4. REFERENCES

[1] S. Haykin, Neural Networks: A Comprehensive Foundation, MacMillan College Publishing Company, New York, 1995.

[2] Kanellopoulos, I., Varfiss, A., Wilkinson, G.G. and Megier, J., 1991: Neural network classification of multi-data satelite imagery. Proceedings of International Geoscience and Remote Sensing Symposium (IGARSS'91), Espoo, Finland, June, pp. 2215-2218.

[3] Rich Drewes "An artificial neural network spam classifier", Project home page: www.interstice.com/drewes/cs676/spam-nn

[4] Sexton, S.R., Dorsey, R.E., "Reliable classification using neural networks: a genetic algorithm and back-propagation comparison", ***Decision Support Systems***, 30: 11–22 (2000).

[5] Rumelhart, D.E., Hinton, G., Williams, R., "Learning representation by back-propagation errors", ***Nature***, 323(9): 533-536 (1986).

[6] W. J. J. Rey, "Introduction to Robust and Quasi-Robust Statistical Methods", Springer-Verlag, Berlin, 110–116, 1983.

[7] J. Yu, J. Amores, N. Sebe, Q. Tian, "Toward an improve Error metric", International Conference on Image Processing (ICIP), 2004.

[8] J. Yu, J. Amores, N. Sebe, Q. Tian, "Toward Robust Distance Metric Analysis for Similarity Estimation", IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2006.

[9] [9] R. Battiti, "First and second-order methods for learning: between steepest descent and Newton's method", Neural Computation, Vol. 2, 141-166, 1992.

Group	Content	Example of Keywords in Each Group
C1	Adult	adult, aphrodisiac, big, cam, climax, company, cum, desire, erotic, fantasy, fuck, gay, girl, greate, guy, hard, hardcore, heaven, hot, huge, long, man, max, maxlength, nude, orgasm, penis, performance, pheromone, pill, porn, powerful, pussy, satisfy, sex, stamina, sweet, teen, viagra, webcam, x, xxx, xxx-porn, young, love, teen, anus
C2	Financial	Account, accountant, alert, analyst, attorney, bank, bankruptcy, benefit, bill, billing, broker, budget, building, cash, cheque, commission, consolidate, court, credit, creditor, currency, customer, debt, deposit, discover, economy, entrepreneur, estate, exchange, fee, finance, freedom, fund, help, high-risk, insurance, invest, investor, judgment, legal, legitimate, lender, loan, mastercard, mortgage, obligate, pay, payable, payable, paycheck, promote, purchase, rate, refinance, refund, rent, revenue, risk, service, statement, stock, support, tax, transaction, vat, visa, wealth, worth, service
C3	Commercial	college, commerce, computer, cost, deliver, discount, especial, expensive, express, fantastic, free, furnishing, furniture, game, get, gif, gift, great, guarantee, inexpensive, invite, item, just, keyboard, license, lifetime, magazine, maintenance, mall, market, material, materials, mobile, motherboard, mouse, offer, online, only, order, palm, pamphlet, percent, premium, price, produce, product, program, recommend, refill, release, resell, reseller, retail, sale, save, save, sell, ship, shipping, shop, shopping, special, subscribe, supply, surprise, trade, trademark, upgrade, voucher, whole, wholesale, within
C4	Beauty and Diet	after, age, amaze, anti-aging, appetite, beauty, become, before, believe, blood, body, botanic, breast, build, burn, Diet calorie, capsule, card, cell, change, chemical, cholesterol, confirm, course, diet, difference, dose, drug, effect, effective, eliminate, energy, enhance, exercise, eye, face, fast, fat, firm, fit, fitness, flexible, gary, grow, grown, growth, hair, health, healthcare, heart, height, herb, herbal, hormone, improve, inche, incredible, kidney, large, laser, life-changing, light, lose, loss, low, magic, medicine, metabolism, micro-cap, miracle, modem, move, muscle, nature, nutrient, old, over, overweight, permanent, plain, potential, pound, power, protect, reduce, remanufacture, repair, restore, retain, reverse, safe, satisfaction, secret, size, step, strength, strong, tablet, therapy, thin, toxin, treatment, under, virginia, vitamin, weight, woman, wonderful, wrinkle
C5	Traveling	book, deluxe, excite, guide, holiday, honest, hotel, luxury, meal, package, plan, problem, relax, relief, reserve, resort, summer, temple, ticket, tour, train, travel, traveler, trip, vacation,
C 6	Home-Based	address, astonishment, base, broadcast, bulk, business, comfort, connect, demo, domain, downline, download, Business earn, email, emailing, ethernet, facemail, fresh, home, homebased, homeworker, host, income, interest, international, internet, investigate, job, list, lucrative, mail, mailbox, mailer, mailing, make, marketing, message, million, money-making, opportunity, part-time, people, private, profit, reach, receive, recipient, require, re-register, return, server, software, subscriber, success, teach, unsubscribe, user, visit, website, work, work-at-home, worker, working
C7	Gambling	action, award, bet, bonus, casino, challenge, extra, gambling, gold, hunt, las, lucky, millionaire, player, poker, prize, reward, rich, vegas, win, lottery