Fuzzy Output Error as the Performance Function for Training Artificial Neural Networks to Predict Reading Comprehension from Eye Gaze

Leana Copeland, Tom Gedeon, and Sumudu Mendis

Research School of Computer Science
Australian National University
Canberra, Australia
{leana.copeland,tom.gedeon,sumudu.mendis}@anu.edu.au

Abstract. Imbalanced data sets are common in real life and can have a negative effect on classifier performance. We propose using fuzzy output error (FOE) as an alternative performance function to mean square error (MSE) for training feed forward neural networks to overcome this problem. The imbalanced data sets we use are eye gaze data recorded from reading and answering a tutorial and quiz. The goal is to predict the quiz scores for each tutorial page. We show that the use of FOE as the performance function for training neural networks provides significantly better classification of eye movements to reading comprehension scores. A neural network with three hidden layers of neurons gave the best classification results especially when FOE was used as the performance function for training. In these cases, upwards of a 19% reduction in misclassification was achieved compared to using MSE as the performance function.

Keywords: Eye tracking, reading comprehension prediction, fuzzy output error (FOE), imbalanced data sets, performance function.

1 Introduction

In this analysis we look at the practical application of predicting reading comprehension based on eye gaze recorded from participants while they read and completed a quiz. We have found no published papers on predicting reading comprehension using artificial neural networks. Current applications of eye tracking in reading analysis only take into account basic assessment of reading behavior such as using fixation time to predict when a user pauses on a word. We intend to explore the use of more complex analysis of eye gaze to make more complex prediction about the users reading behavior. We do this by investigating the use of artificial neural networks to predict these complex behaviors. However, this application poses us with several obstacles namely restricted size in the data sets that are highly imbalanced. We explore a method for improving classification performance of artificial neural networks (ANN) in this scenario. We investigate the use of fuzzy output error (FOE) [1] as the performance function for training the feed forward neural networks using back propagation training.

C.K. Loo et al. (Eds.): ICONIP 2014, Part I, LNCS 8834, pp. 586–593, 2014.

We assess whether the use of this performance measure is better suited to this type of problem compared to mean square error (MSE).

The intended use of reading comprehension prediction from eye gaze is in the design of adaptive online learning environments that use eye gaze to predict user reading behavior.

2 Background

2.1 Eye Movements during Reading

Eye movements can be broadly characterized as fixations and saccades. A fixation is where the eye remains relatively still to take in visual information. A saccade is a rapid movement that transports the eye to another fixation. Generally when reading English fixation duration is around 200-300 milliseconds, with a range of 100-500 milliseconds and saccadic movement is between 1 and 15 characters with an average of 7-9 characters [2]. The majority of saccades are to transport the eye forward in the text when reading English, however, a proficient reader exhibits backward saccades to previously read words or lines about 10-15% of the time [2]. Backward saccades are termed regressions. Long regressions occur due to comprehension difficulties, as the reader tends to send their eyes back to the part of the text that caused the difficulty [2]. Comprehension of the text can have significant effects on the eye movements observed [2,3]. Eye gaze patterns can be used to differentiate when individuals are reading different types of content [4]. In this application both support vector machines (SVM) and ANN were used to classify eye movement measures as either relevant or irrelevant text for answering a set of questions. ANN's have also been used to predict item difficulty in multiple choice reading comprehension tests [5]. Their analysis took into account the text structure, propositional analysis of the text, and the cognitive demand of the text, but not eye gaze.

2.2 Performance Functions for Imbalanced Data Sets

Dealing with imbalanced data sets is not a new problem. Performance functions for dealing with imbalance in data sets include increasing the weight-updating for the minority class and decreasing it for the majority class [6,7]. This error function was designed specifically for use in the back-propagation algorithm for training feed forward artificial neural networks. Many other methods have been used to overcome the problem of imbalanced data sets such as using under-sampling, over-sampling, and other forms of sampling to reduce the imbalance. An example of a cost sensitive learning algorithm is MetaCost [8] which is based on relabelling of training data with their estimated minimal cost classes. Another way of achieving cost sensitivity is to change the algorithm used to train the classifier to utilize a cost matrix, such as with neural networks [9,10].

2.3 Fuzzy Output Error (FOE)

Fuzzy Output Error (FOE) [1] is an extension of Fuzzy Classification Error (FYCLE) and Sum of Fuzzy Classification Error (SYCLE) [11]. However, FOE uses a fuzzy membership function to measure the difference between the predicted and the target values. Instead of mean square error (MSE), FOE describes the error in a fuzzy way and then sums the fuzzy errors to get the total error. FOE is defined as follows for a data set of n records with matching pairs of target and predicted values for each record 1 to n: $FOE = \sum_{i=1}^{n} 1 - \mu(\hat{y}_i - y_i)\ where\ n \in \mathbb{N}$ and $\mu()$ is the membership function of a desired classification and its complement describes the error. The membership function is termed the FOE Membership Function, which we will refer to as FMF subsequently.

The FMF is used to describe the output of a fuzzy classification (or a regression) in regards to how close that output is to the target output. The membership function itself represents the fuzzy set for "good classification". The value of $\mu(x)$ gives the degree of membership of the error in the good classification fuzzy set and consequently the complement of $\mu(x)$ gives the error measure. In the case of perfect classification $\hat{y} - y = 0$ so the membership value is $\mu(\hat{y} - y) = 1$. Conversely, when $\hat{y} - y = 1$ the classification is completely wrong so the membership value is $\mu(\hat{y} - y) = 0$. FOE can represent crisp classification, i.e. the special case of $\mu(x) \in \{1,0\}$. The more $\mu(x)$ tends toward 0 the higher the error, since the difference is larger. FMFs can be created in any shape in order to describe the output of a function. It is important to note that the difference between target and predicted values is not taken as the absolute value of the difference (i.e. $|\hat{y} - y|$). Although this would make the FMF simpler as only one side of a piecewise linear function would be needed, it provides more flexibility in describing the types of error. For example, false negatives may be considered a much worse error than false positives when screening for diseases.

3 Method

A user study was conducted to collect participants' eye gaze as they read a tutorial and completed a quiz based on the tutorial's content. The tutorial and quiz were coursework from a first year computer science course at the Australian National University. The tutorial and quiz were presented to participants in two formats. The first format (denoted by A) involved presentation of the tutorial content slide followed by questions and the content slide. As there are 9 topics there are 18 slides in total displayed in this format. The second format (B) involved presentation of the questions and the content slide and so there are 9 slides in total displayed in this format. Each of the 9 slides is 400 words long with an average Flesch Kincaid Grade Level[1] of 12. All participants were university students and therefore had at least high school level education indicating that the readability of the slides should not be above their reading

[1] Flesch Kincaid Grade Level is an indication of the minimum level of education required to read and comprehend a piece of text. The Flesch Kincaid readability test is designed for contemporary English and United States educational system grading.

abilities. Participants answered two questions to measure their comprehension (18 questions in total); one question is multiple-choice and the other is cloze (fill-in-the-blanks). The two types of questions are to assess different forms of comprehension. The scores that the participants can receive for each question are 0, 0.5 and 1. Once the participants finished the quiz and before being shown their results, participants were asked to subjectively rate their overall comprehension on a scale of 1 to 10 with 10 being complete understanding.

Format A was presented to 15 participants (6 female, 9 male) with an average age of 22.3 years. Of these participants, 7 stated that their degree or major was related to computer science or information technology. English was not the first language for 4 participants. Format B was presented to 8 participants (1 female, 7 males), with an average age of 21.8 years. All participants stated that they had a major or degree related to computer science. English was not the first language for 3 participants.

The study was displayed on a 1280x1024 pixel monitor. Eye gaze data was recorded at 60Hz using Seeing Machines FaceLAB 5 infrared cameras mounted at the base of the monitor. The study involved a 9-point calibration sequence. EyeWorks Analyze was used to pre-process the gaze point data to give fixation points. The parameters used for this were a minimum duration of 0.06 seconds and a threshold of 5 pixels.

3.1 FMF Shapes Used to Calculate FOE

In this analysis we investigated one FMF shape used for calculating FOE. This FMF (Fig. 1) is designed to be a model of FYCLE.

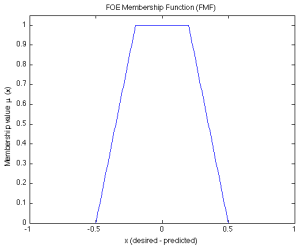

Fig. 1. FMF used to calculate FOE

3.2 Data Set Information

The raw eye gaze data consists of x,y-coordinates recorded at equal time samples (60Hz). Beyond fixation and saccade identification many other eye movement measures can be derived that reveal much about the participants' reading behavior such as; maximum fixation duration (seconds), average fixation duration (seconds), total fixation duration (seconds), and regression ratio. The number of inputs varies depending on the presentation method as the inputs are generated from the pages that the participant viewed. This means that in format A as the participants view the tutorial content page and then the questions and content page, the inputs are generated from both pag-

es for the scores obtained from the questions and content page. Since there is a large difference in the ranges for each of the inputs we normalized the inputs to a range of [0,1]. The two outputs for all data sets are the multiple choice question score and cloze question score. The multiple-choice score can take values of 0 or 1 and the cloze score can take the values 0, 0.5 or 1. This is therefore a classification problem; a binary classification task for the multiple-choice score and a 3-class classification task for the cloze score. However, as shown in Table 1 the ratio of the number of data instances in each class for each problem is considerably imbalanced for each output.

Table 1. Properties of each data set

Properties of data set	**Format A**	**Format B**
Number of Inputs	49	26
Size	135	72
Multiple choice score class imbalance	109/26	59/13
Percentages in classes 1/0	*81%/19%*	*82%/18%*
Cloze Score class imbalance	124/11/0	69/1/2
Percentages in classes 1/0.5/0	*92%/8%/0%*	*96%/1%/3%*

4 Results and Discussion

Several ANN architectures were trained using the scaled conjugate gradient algorithm [12] and FOE used as the performance function. As a comparison the same ANN architectures were trained using MSE as the performance function and the Levenberg–Marquardt algorithm [13]. The number of inputs for each presentation format is outlined in Table 1 and all networks have 2 outputs. From initial testing it was found that a single layer network performed poorly with average misclassification rate (MCR) around the 0.5 for all both FOE and MSE. We have chosen two and three layer topologies to trial for the analysis. The following topologies were tested: [10 5], [20 10], [30 15], [50 25], [12 6 3], [16 8 4], [20 10 5], [30 20 10], and [60 30 15]. The notation [X Y Z] indicates neurons in the first hidden layer to the third hidden layer. As a baseline comparison MSE is used as one of the performance functions. Reported are the average misclassification rate (MCR) values from 10-fold cross validation with standard deviations, summarized in Table 2 and Table 3.

For format A, on average the MCR produced from using FOE as the performance function for training the neural networks to predict the question scores is lower than that from using MSE as the performance function. However, the results are not statistically different. However, there is a statistically significantly difference between the mean MCR values from 10-fold cross validation for each topology for format B ($p=0.02<0.05$, 2-sided, paired Student's t-test). Therefore, on average the MCR produced from using FOE, as the performance function for training the neural networks to predict the question scores is lower than that from using MSE as the performance function.

Table 2. Comparison of MCR from using FOE and MSE as the performance function for training ANNs to classify the Format A data set

Topology	MCR				Difference in MCR	% Reduction in MCR
	FOE Result		MSE Result			
	Mean	*St. Dev.*	*Mean*	*St. Dev.*		
[10 5]	0.32	0.29	0.25	0.10	-0.06	-25.3
[20 10]	0.42	0.28	0.35	0.09	-0.07	-19.8
[30 15]	0.33	0.16	0.38	0.15	0.04	11.9
[50 25]	0.28	0.06	0.33	0.12	0.05	14.2
[12 6 3]	0.23	0.25	0.23	0.14	0.00	-0.4
[16 8 4]	0.20	0.08	0.27	0.07	0.07	24.5
[20 10 5]	0.21	0.06	0.26	0.15	0.05	19.2
[30 20 10]	0.24	0.05	0.31	0.11	0.07	22.8
[60 30 15]	0.25	0.08	0.39	0.13	0.13	34.8
Average	*0.28*	*0.15*	*0.31*	*0.12*	*0.03*	*9.11*

Table 3. Comparison of MCR from using FOE and MSE as the performance function for training ANNs to classify the Format B data set

Topology	MCR				Difference in MCR	% Reduction in MCR
	FOE Result		MSE Result			
	Mean	*St. Dev.*	*Mean*	*St. Dev.*		
[10 5]	0.30	0.26	0.29	0.14	0.00	-1.5
[20 10]	0.27	0.15	0.37	0.15	0.10	26.7
[30 15]	0.33	0.21	0.46	0.13	0.13	29.1
[50 25]	0.52	0.22	0.47	0.18	-0.04	-9.3
[12 6 3]	0.16	0.07	0.24	0.13	0.07	31.1
[16 8 4]	0.16	0.06	0.30	0.14	0.14	46.4
[20 10 5]	0.16	0.07	0.24	0.10	0.07	30.6
[30 20 10]	0.26	0.11	0.39	0.15	0.13	32.6
[60 30 15]	0.35	0.22	0.36	0.10	0.01	2.5
Average	*0.28*	*0.15*	*0.35*	*0.14*	*0.07*	*20.91*

Overall, the results reflect that fact that the data sets are quite hard to classify. This could be due to several factors: class imbalance, small data sets, and too many feature inputs. However, these obstacles can be common in real world problems so it is imperative that such obstacles can be overcome. FOE has been shown to be a flexible performance function that can be tailored specifically for each problem. By defining different error membership functions (the FMF used) for FOE the outcome of training ANNs to classify the eye movement measures can be improved compared to using MSE. This is shown for both data sets where on average the use of FOE as the performance function for training the neural networks produces neural networks that are better as predicting the multiple choice and cloze scores.

Notably, for both data sets the topologies that generate the best predictions are [16 8 4], [20 10 5], and [30 20 10]. This reiterates the fact that the data set is hard to classify and contains complex relationships, as three layers of hidden neurons are needed to provide decent classification results. Furthermore, using FOE as the performance function for training generates upwards of a 19% reduction in misclassification compared to using MSE. Particularly, when using the [16 8 4] topology and FOE as the performance function for training creates a neural network that produces on average 38% and 46% reduction in misclassification, for formats A and B respectively, compared with using MSE.

5 Conclusions and Further Work

We have shown that the use of FOE as a performance function for training feed forward neural networks provides better classification of results than using MSE when the data is imbalanced. The use of FOE as the performance function for training neural networks provides significantly better classification of eye movements than reading comprehension scores. We found that the eye movement data is quite complex so it is optimal to use a neural network with three hidden layers of neurons. In these cases the use of FOE as the performance function for training gave upwards of a 19% reduction in misclassification compared to using MSE as the performance function, with a maximum of 46% reduction in misclassification, which is a significant improvement in classification. These are promising results and show that when dealing with a small data set with a large imbalance in classes MSE is not the optimal performance function to use for training neural networks. Further work will be needed to generalize to other data sets as well as with other classifiers. Additionally, we intend to extend this analysis to compare to existing techniques for handling imbalanced data sets such sampling methods and cost-sensitive learning.

One of the advantages of using FOE is that it is a flexible error function that can be tailored to the data sets and problem. Specifying the shape of the FMF used to calculate FOE does this. However, there is no simple way of constructing an FMF. In this analysis we only investigated one FMF. Other FMF shapes should be tested such as those described in [14]. However, a beneficial approach would be to learn the most appropriate FMF shape from the data set. An initial investigation on how to do this was also done in previous work but was restricted to looking at fuzzy signatures [14]. An area of further exploration is how to apply the learning of FMF shape when using other classifiers such as neural networks.

The application of predicting reading comprehension from eye gaze is in adaptive online learning environments. Prediction of comprehension would allow a system to adaptively change to a student's knowledge level making the learning process more streamlined and targeted toward their capabilities. Much is left to do in this respect. A primary area of interest is in predicting reading comprehension without questions. In both scenarios here the participants had access to the questions and the tutorial content at the same time so that they could cross-reference the text and questions to find the most appropriate answer. In a scenario where the student is shown text and no

comprehension questions it would be beneficial to be able to predict their comprehension without needing to interrupt them with comprehension questions.

References

1. Gedeon, T., Copeland, L., Mendis, B.S.U.: Fuzzy Output Error. Australian Journal of Intelligent Information Processing Systems 13(2), 37–43 (2012)
2. Rayner, K.: Eye movements in reading and information processing: 20 years of research. Psychological Bulletin, 372–422 (1998)
3. Rayner, K., Chace, K.H., Slattery, T.J., Ashby, J.: Eye movements as reflections of comprehension processes in reading. Scientific Studies of Reading 10(3), 241–255 (2006)
4. Vo, T., Mendis, B.S.U., Gedeon, T.: Gaze Patterns and Reading Comprehension. In: Wong, K.W., Mendis, B.S.U., Bouzerdoum, A. (eds.) ICONIP 2010, Part II. LNCS, vol. 6444, pp. 124–131. Springer, Heidelberg (2010)
5. Perkins, K., Gupta, L., Tammana, R.: Predicting item difficulty in a reading comprehension test with an artificial neural network. Language Testing 12(1), 34–53 (1995)
6. Oh, S.-H.: Error back-propagation algorithm for classification of imbalanced data. Neurocomputing 74(6), 1058–1061 (2011)
7. Oh, S.-H.: Improving the Error Back-Propagation Algorithm for Imbalanced Data Sets. International Journal of Contents 8(2), 7–12 (2012)
8. Domingos, P.: MetaCost: a general method for making classifiers cost-sensitive. In: Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 155–164. ACM (1999)
9. Kukar, M., Kononenko, I.: Cost-Sensitive Learning with Neural Networks. In: 13th European Conference on Artificial Intelligence, pp. 445–449 (1998)
10. He, H., Garcia, E.A.: Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering 21(9), 1263–1284 (2009)
11. Mendis, B.S.U., Gedeon, T.D.: A comparison: Fuzzy signatures and Choquet Integral. In: IEEE International Conference on Fuzzy Systems, FUZZ-IEEE 2008 (IEEE World Congress on Computational Intelligence), pp. 1464–1471 (2008)
12. Moller, M.F.: A scaled conjugate gradient algorithm for fast supervised learning. Neural Networks 6(4), 525–533 (1993)
13. Hagan, M.T., Menhaj, M.: Training feedforward networks with the Marquardt algorithm. IEEE Transactions on Neural Networks 5(6), 989–993 (1994)
14. Copeland, L., Gedeon, T.D., Mendis, B.S.U.: An Investigation of Fuzzy Output Error as an Error Function for Optimisation of Fuzzy Signature Parameters. RCSC TR-1 2014 (2014)