Using Misclassification Analysis for Data Cleaning

Piyasak Jeatrakul, Kok Wai Wong, and Chun Che Fung
School of Information Technology, Murdoch University
Western Australia 6150
Email: [p.jeatrakul | k.wong | l.fung] @murdoch.edu.au

Data cleaning is a pre-processing technique used in most data mining problems. The purpose of data cleaning is to remove noise, inconsistent data and errors in order to obtain a better and representative data set to develop a reliable prediction model. In most prediction model, unclean data could sometime affect the prediction accuracies of a model. In this paper, we investigate classification problem, which make use of misclassification analysis technique for data cleaning. To demonstrate our concept, we have used artificial neural network (ANN) as the core computational intelligence technique. We use three benchmark data sets obtained from the University of California Irvine (UCI) machine learning repository to investigate the results from our proposed data cleaning technique. The experimental data sets used in our experiment are binary classification problems, which are German credit data, BUPA liver disorders, and Johns Hopkins Ionosphere. The results from our experiments show that the proposed cleaning technique could be a good alternative to provide some confidence when constructing a classification model.

Keywords: Data cleaning, data pre-processing, artificial neural network, classifier.

1. Introduction

When constructing a prediction model, regardless it is for classification or function approximation, it is always difficult to have an exact function that describes the relationship between the input vector, *X* and target vector, *Y*. However, a probabilistic relationship govern by joint probability law ν can be used to describe the relative frequency of occurrence of vector pair (X_n,Y_n) for *n* training set. The joint probability law ν can further separate into environmental probability law μ and conditional probability law γ. For notation expression, the probability law can be expressed as:

$$P(\nu) = P(\mu)P(\gamma)$$

For environmental probability law μ, it describes the occurrence of *X*. As for conditional probability law γ, it describes the occurrence of *Y* given *X*. A vector pair *(X, Y)* is considered as noise if *X* does not follow the environmental probability law μ, or the *Y* given *X* does not follow the conditional probability law γ.

With the assumption that most data sets have noise, it is desirable to provide some confidence in detecting the noise in the data set. After we can identify the noise confidently, we can then remove them from the training data set. In this paper, we focus on classification problems. We present our proposed misclassification analysis technique suitable for data cleaning. The core techniques used in our study is based on artificial neural networks (ANNs).

Classification is a process when an object needs to be classified into a predefined class or group based on attributes of that object. There are many real world applications that can be categorized as classification problems such as weather forecast, credit risk evaluation, medical diagnosis, bankruptcy prediction, speech recognition, handwritten character recognition, quality control, and engineering [1], [2], [3].

In classification problems, the data pre-processing is a significant process before developing the classification model. Generally, a data set may consist of some undesirable data used to develop the classification model. As mentioned before, it may consist of noise and inconsistent data. Therefore, pre-processing techniques are normally used to enhance the data used for establishing the classification model [4].

In recent years, there are several studies in pre-processing techniques related to classification problems. For example, Tang et al [5] used the morphological data cleaning algorithms to deal with noisy data. This technique effectively improves the classification performance when comparing with methods used in their comparison. Brodley et al [6] used a set of learning algorithms to identify and eliminate noise before developing the classification model in order to improve the classification accuracy. This technique can reduce the noise level by up to 30%. Kubica et al [7] proposed an iterative and probabilistic approach to identify and remove corrupted data from the training data. This approach

can improve the overall classification accuracy. Furthermore, Setiawan et al [8] applied pre-processing for the heart disease database by removing the records that have too many missing values and removing outliers using statistical methods. They also combined Artificial Neural Network (ANN) with rough set theory (RST), named as ANNRST, to predict some missing values. Moreover, Zhimin et al [9] proposed a new framework dealing with missing value. They used back-propagation neural network to predict missing value. In addition, they used Adaptive Boosting (AdaBoost) to classify data. The experimental results show that the accuracy of classification increases significantly with their algorithms.

Most of the researches try to increase the quality of training data by using some form of pre-processing, so as to increase the classification accuracy. However, majority of the researchers attempt to deal with missing data and outlier data. It can be observed that not many of the researcher performed further analysis on those misclassification patterns. This is thus the purpose of this paper to move one step forward in misclassification analysis to improve the classification accuracy. There is suggestion that by understanding the nature of misclassification data, the classification accuracy may be able to be improved [10].

In this paper, we formulate a technique to perform misclassification analysis in the wish that we can identify noisy data with some confidence. After identifying the noisy data, we can then perform data cleaning. In this paper, we apply the concept from the Complementary Neural Network (CMTNN) [11] as the cleaning technique to enhance the performance of a neural network classifier. CMTNN is selected because of its particular characteristics. It can integrate the truth and false membership values to deal with the uncertainty in classification while other techniques use only truth membership values.

In the experiments, three binary classification data sets from the University of California Irvine (UCI) machine learning repository [12] are used. These include German credit data, BUPA liver disorders, and Johns Hopkins Ionosphere. These data sets are selected because they are benchmark data sets which have been commonly used in the literature.

2. Cleaning Techniques Using Misclassification Analysis

In this section, the concept of Complementary Neural Network (CMTNN) is described. The proposed cleaning techniques based on CMTNN will then be presented.

2.1 Complementary Neural Network (CMTNN)

CMTNN [11] is a technique using a pair of complementary ANNs called Truth neural network and Falsity neural network as shown in Fig 1.

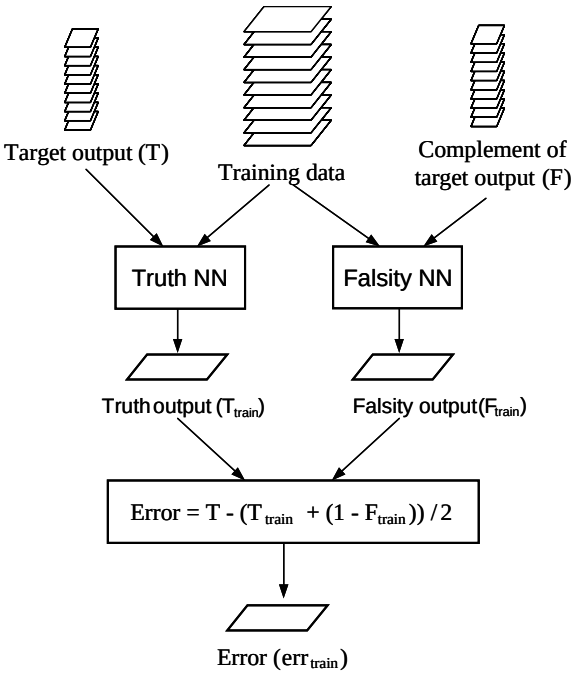

Figure 1: Complementary neural network

This technique has successfully been implemented for both binary and multiclass classification problems. For binary classification, a pair of neural networks is implemented in order to predict degrees of truth and false membership values. The predicted results of Truth NN and Falsity NN are compared to provide the classification outcomes. The difference between the truth and false membership values can also be used to represent uncertainty in the classification [11],[13].

2.2 Cleaning Techniques

In order to apply CMTNN for data cleaning, Truth NN and Falsity NN are employed to detect the misclassification patterns. There are basically two ways that we can perform the cleaning. For the purpose of this paper, we will include these two possible cleaning techniques for discussion. The techniques are used to discover and clean misclassification patterns from a training set. In this case, we assume that the misclassified patterns as noisy data. The steps of these two cleaning techniques are described as follows.

Cleaning Technique I:

a. Prepare the training data for Falsity NN by complementing the target outputs of the training set.

b. The Truth NN and Falsity NN are trained by truth and false membership values.

c. The prediction outputs on the training sets of both NNs are compared with the actual outputs. The misclassification patterns are also detected if the prediction outputs and actual outputs are different.

d. In the last step, the new training set is cleaned by eliminating all misclassification patterns detected by the Truth NN (T_{mis}) and Falsity NN (F_{mis}) respectively, i.e. $T_{mis} \cup F_{mis}$. As for training a new neural network classifier, the cleaned data set that removes those misclassification patterns will be used. Please take note that if the misclassification patterns appear in both Truth NN and Falsity NN, i.e. duplication, only one set will be used.

Cleaning Technique II:

a. Repeat the step *a.* to *c.* of cleaning technique I.

b. The new training set is cleaned by eliminating only the misclassification patterns detected by both the Truth NN (T_{mis}) and Falsity NN (F_{mis}), i.e $T_{mis} \cap F_{mis}$.

3. Experiments and Results

Three data sets from UCI machine learning repository [12] are used in the experiment. The data sets for binary classification problems include German credit data, BUPA liver disorders, and Johns Hopkins Ionosphere.

- The purpose of German credit data set is to predict whether a loan application is "Good" or "Bad" credit risk.
- The purpose of BUPA liver disorders data set is to predict whether a male patient shows signs of liver disorders.
- The purpose of Johns Hopkins Ionosphere data set is to predict "Good" or "Bad" radar return from the ionosphere.

The characteristics of these three data sets are shown in Table 1.

Table 1. Characteristics of data sets used in the experiment.

Name of data set	No. of patterns	No. of attributes	No. of patterns in class 1	No. of patterns in class 2
German credit data	1000	20	700	300
BUPA liver disorders	345	6	145	200
Johns Hopkins Ionosphere	351	34	225	126

For the purpose of establishing the classification model and testing it, each data set is first split into 80% training set and 20% test set as shown in Table 2. Furthermore, the cross validation method is used to obtain reasonable results. Each data set will be randomly split ten times to form different training and test data sets. For the purpose of this study, the results of the ten experiments of each data set will be averaged.

Table 2. Number of patterns in the training and test sets.

Name of data set	No. of training data	No. of test data	Total
German credit data	800	200	1000
BUPA liver disorders	276	69	345
Johns Hopkins Ionosphere	281	70	351

Table 3 shows the average number of misclassification patterns in each data set detected by Truth NN and Falsity NN. The results show that the number of misclassification patterns detected by both NNs is almost similar. For example, in German credit data, misclassification patterns detected by Truth NN and Falsity NN are 169 and 165 patterns respectively. Furthermore, there are also misclassification patterns discovered by both NNs, i.e. the same patterns that are misclassified by Truth NN as well as the Falsity NN. They are 125, 55 and 6 such patterns for German credit, BUPA liver disorders, and John Hopkins Ionosphere data set respectively.

After the training sets are cleaned by the two proposed cleaning techniques as mentioned in section 2, new neural network classifiers are trained by the cleaned training sets. The performance of each

classifier for the training set and test set before and after cleaning data are evaluated. The comparison results are shown in Table 4.

Table 3. Average number of misclassification patterns of the training sets.

Name of data set	No. of misclassification patterns detected by Truth NN	No. of misclassification patterns detected by Falsity NN	No. of the misclassification patterns detected by both NNs
German credit data	169	165	125
BUPA liver disorders	79	77	55
Johns Hopkins Ionosphere	10	7	6

Table 4. Average classification accuracy (%) of the test sets before and after cleaning data.

Name of data set	Before cleaning	After cleaning training data with technique I	After cleaning training data with technique II
German credit data	76.25	76.95	77.55
BUPA liver disorders	69.99	70.14	71.01
Johns Hopkins Ionosphere	90.29	91.71	92

From the comparison results in Table 4, we found that cleaning technique II can increase the classification performance on the test sets better than cleaning technique I. While, the classification accuracies using cleaning technique I increase from 76.25% to 76.95% on German credit data, from 69.99% to 70.14% on BUPA liver disorders data, and from 90.29% to 91.71% on Johns Hopkins Ionosphere. On the other hand, the performance after cleaning with technique II on each data set increase to 77.55%, 71.01% and 92% on German credit data, BUPA liver disorders, and Johns Hopkins Ionosphere respectively. Cleaning technique II performs well on the test sets because this technique removes only the high potential misclassification patterns rather than eliminates all possible misclassification patterns from the training set. This also suggested that cleaning technique II can provide more confidence in noise identification.

Although the improvement of the accuracies in this case study may not be significant, the proposed technique is able to provide a mean to increase the confidence of identifying the noisy data. This can be viewed as a first step in the proposed misclassification analysis used for data cleaning. There are also many factors that can be optimized in future to study the behaviour of the proposed misclassification analysis. The proportion of separating the training and testing data may be re-distributed to investigate the distribution of the training and testing set. Another danger for cleaning the noisy data is overtraining the ANN, more rigid generalization techniques could be experiment to study the behaviour of the model after the noisy data have been removed.

4. Conclusions

This paper presents the proposed misclassification technique to increase the confidence of cleaning noisy data used for training. In this paper, we focus our study for classification problem using ANN. The CMTNN is applied to detect misclassification patterns. In the experiment, the training data is cleaned by two cleaning techniques. For technique I, the training data is cleaned by eliminating all misclassification patterns discovered by the Truth NN and Falsity NN. For technique II, training data is cleaned by eliminating only the misclassification patterns discovered by both the Truth NN and Falsity NN. After misclassification patterns are removed from the training set, a neural network classifier is trained using the cleaned data. In the experiment of this paper, three data sets from the University of California Irvine (UCI) machine learning repository including German credit data, BUPA liver disorders, and Johns Hopkins Ionosphere are used. The neural network classifiers have also been evaluated and compared in terms of their performances. Results obtained from the experiment indicated this initial study could be carried further to optimize the misclassification analysis to be used as an alternative to improve prediction model.

5. References

1 P. Kraipeerapun, C. C. Fung, W. Brown, K. W. Wong, and T. Gedeon, "Uncertainty in mineral prospectivity prediction," in *the 13th International Conference on Neural Information Processing ICONIP 2006*, Hong Kong, 2006, pp. 841-849.

2 G. P. Zhang, "Neural networks for classification: a survey," *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, vol. 30, pp. 451-462, 2000.

3 P. M. Wong, I. J. Taggart, and T. D. Gedeon, "The use of neural network methods in porosity and

permeability predictions of a petroleum reservoir," *AI Applications*, vol. 9(2), pp. 27-38, 1995.

4 H. Frigui, "Pre-processing for data clustering," in *Fuzzy Information, 2004. Processing NAFIPS '04. IEEE Annual Meeting of the*, 2004, pp. 967-972 Vol.2.

5 S. Tang and S.-p. Chen, "Data cleansing based on mathematic morphology," in *Bioinformatics and Biomedical Engineering, 2008. ICBBE 2008. The 2nd International Conference on*, 2008, pp. 755-758.

6 C. E. Brodley and M. A. Friedl, "Identifying mislabeled training data," *Journal of Artificial Intelligence Research*, vol. 11, pp. 137-167, 1999.

7 J. Kubica and A. Moore, "Probabilistic noise identification and data cleaning," in *Data Mining, 2003. ICDM 2003. Third IEEE International Conference on*, 2003, pp. 131-138.

8 N. A. Setiawan, P. Venkatachalam, and A. F. M. Hani, "Missing Attribute Value Prediction Based on Artificial Neural Network and Rough Set Theory," in *BioMedical Engineering and Informatics, 2008. BMEI 2008. International Conference on*, 2008, pp. 306-310.

9 M. Zhimin, P. Zhisong, H. Guyu, and Z. Luwen, "Treating missing data processing based on neural network and AdaBoost," in *Grey Systems and Intelligent Services, 2007. GSIS 2007. IEEE International Conference on*, 2007, pp. 1107-1111.

10 M. Ciraco, M. Rogalewski, and G. Weiss, "Improving classifier utility by altering the misclassification cost ratio," in *Proceedings of the 1st international workshop on Utility-based data mining* Chicago, Illinois: ACM, 2005.

11 P. Kraipeerapun, C. C. Fung, and S. Nakkrasae, "Porosity prediction using bagging of complementary neural networks," in *Advances in Neural Networks – ISNN 2009*, 2009, pp. 175-184.

12 A. Asuncion and D. J. Newman, "UCI Machine Learning Repository," University of California, Irvine, School of Information and Computer Sciences, 2007.

13 P. Kraipeerapun and C. C. Fung, "Comparing performance of interval neutrosophic sets and neural networks with support vector machines for binary classification problems," in *Digital Ecosystems and Technologies, 2008. DEST 2008. 2nd IEEE International Conference on*, 2008, pp. 34-37.