Artificial neural network — Naïve bayes fusion for solving classification problem of imbalanced dataset

Article · April 2011

DOI: 10.1109/ICMSAO.2011.5775584

4 authors, including:

Asrul Adam
University of Malaya
23 PUBLICATIONS **74** CITATIONS
SEE PROFILE

Zuwairie Ibrahim
Universiti Malaysia Pahang
189 PUBLICATIONS **427** CITATIONS
SEE PROFILE

Artificial Neural Network - Naïve Bayes Fusion for Solving Classification Problem of Imbalanced Dataset

Asrul Adam, Mohd Ibrahim Shapiai, Zuwairie Ibrahim, Marzuki Khalid
Faculty of Electrical Engineering
Universiti Teknologi Malaysia
81300, Skudai, Johor, MALAYSIA
asruldm@fkegraduate.utm.my, ibrahim@fke.utm.my,
zuwairiee@fke.utm.my, marzuki@utm.my

Abstract

Incorporating knowledge from domain expert to a classifier is one of the techniques which require to be considered in solving imbalanced dataset problems. In this study, the proposed technique is a development to extend the process for imbalanced dataset where the individual classification system has already been designed for balanced data set. This paper introduces a methodology and preliminary results which are used to investigate whether the proposed approach is possible to improve a classifier's performance when domain expert is employed to the naïve bayes classifier. Domain expert is an additional knowledge which is produced by expert system (neural network) and then become an additional input to the naïve bayes classifier. By using several benchmark data sets from the UCI Machine Learning Repository, the results of the proposed technique show an improvement as compared to the conventional naïve bayes classifier.

Keywords: Domain Expert; Expert knowledge; Naïve Bayes Classifier; Neural network Classifier; Imbalanced data set; Knowledge Engineering

1. Introduction

In this current society, people in related fields are searching the suitable expert systems such as Neural Network [1], Bayesian Network [2], Fuzzy System [3], Probabilistic Machine [4], Decision Tree [5], and Support Vector Machine [6] towards building a classifier for imbalanced data set. The expert system is widely used in many areas of applications such as computer security [7], biomedical [8], remote-sensing [9], engineering [10], and manufacturing industry [11]. The conventional expert systems are modified at data level or algorithmic level or vice versa. The modification of the existing technique is to ensure the application of the proposed technique is to follow the nature of imbalanced dataset and achieves a better prediction performance. Some modifications at data level are usually related to the pre-processing technique such as up-sampling, oversampling, or SMOTE [12], [13] while modifications at algorithmic level focus more on modifying the internal learning algorithm. All these existing classifiers are categorized in machine learning disciplines and largely remain different expertise when they are applied expert knowledge or incorporating that to the expert system which are also so-called knowledge engineering disciplines.

Dybowski *et al*, [14], has mentioned about the machine learning and the knowledge engineering in engineering field. In general, these two techniques can be combined to develop a decision support system suggested by Atish P. Sincha *et al* [15] in his journal.

In order to design a decision support system for imbalanced data sets, this paper has introduced a proposed approach model from existing decision support system by Atish P. Sincha *et al* [15] which is able to tackle the imbalanced dataset problems. Two different expert systems are used for the proposed decision support system, which are the naïve bayes classifier and neural network classifier. Naïve bayes classifier is the main classifier which is supported by the neural network classifier. The predicted outputs from the neural network become a domain expert knowledge to the main classifier. This process is also called data-driven in knowledge engineering discipline which relates to the expert system. The other type is an expertise-driven where the domain expert emerges from the rules given by an expert or engineer in related fields. Domain expert knowledge can be defined as an additional information from several knowledge sources either by using data-driven or expertise-driven, which can be used to guide a classifier in classification process [16]. In the expertise-driven, several knowledge acquisition techniques are used such as interview, protocol analysis, observation and discussion in focus group. The results from these techniques produce relevant rules or protocols. It acts as an additional knowledge to the main classifier. Fuzzy system is one of the tools that able to incorporate expert knowledge to modify the rules in the system. It has been categorized as expertise-driven in knowledge engineering due to the available rules in the Fuzzy System are modified by expert.

The objective of modifying this model is to improve the conventional naïve bayes classifier's performance by incorporating neural network as an expert system through the data-driven type as in the decision support system. In particular, the modification will try to solve the imbalanced dataset problems. The findings from this study are recorded to see the potential of this approach to model the decision support system in solving imbalanced data sets problems.

2. Imbalanced Dataset and its Problems

Imbalanced dataset can be defined by a collection of dataset or samples that consist of several types of input parameters and outputs (classes), where one of the classes is more significant than the other class e.g. the ratio of minority and the majority class is (3:100). Imbalanced data sets are considered high prospects in many research areas such as medical, engineering, remote-sensing and manufacturing. This study only focuses on two classes ("0" or "1") dataset or so-called binary classification.

The classifier that is designed for imbalanced dataset must have an ability to correctly predict both classes and the priority of the prediction focuses more on the minority class samples. Based on this ability, the measure performance is chosen carefully. Thus the selected measure performance should be able to produce fairly calculated accuracy because it is working in imbalanced nature. It resembles unacceptable result if all of the majority classes have an accuracy prediction of (95%) and at the same time it wrongly predicted all of the minority class (0%) even if the classifier's performance gives a high percentage.

Many researchers agree the problems in imbalanced data sets emerge from the difficulty of the conventional classifiers to predict the minority class due to the bias on the majority class [17], [18]. In other words, the conventional classifier is unable to comprehend the correct output as because the conventional classifiers are mainly designed for balanced data sets. To generalize the proposed approach, this paper has ignored the types of input parameters and only considers the ratio of the two classes (majority and minority classes) for the imbalanced data set.

2.1 Performance Measure

In this study, geometric mean (Gmean) is used as performance measure for the classifier. The Gmean can be defined as follows:

$$Gmean = \sqrt{(TNR \times TPR)} \quad (1)$$

where,

$$TNR = \frac{TN}{TN + FP} \quad (2)$$

$$TPR = \frac{TP}{TP + FN} \quad (3)$$

where, TP, FN, FP, and FN can be defined as follows. True Positive (TP), correctly predicts the majority class (0). False Negative (FN), wrongly predict the minority class to be majority class and False Positive (FP), also wrongly predicts majority class to be minority class. True Negative (TN), correctly predicts minority class (1). Table 1 shows a confusion matrix for outcome of a two class problem. Positive means majority class and negative means minority class. According to [19], this table is usually used as a basis for various measures.

Table 1. Confusion matrix for a 2-class problem

	Predicted Positive	Predicted Negative	
Actual Positive	True Positive (TP)	False Negative (FN)	Total Actual Positive
Actual Negative	False Positive (FP)	True Negative (TN)	Total Actual Negative
	Total Predicted Positive	Total Predicted Negative	Total Units

Table 2. Summary of Benchmark Dataset provided by UCI Machine Learning Repository

Dataset	**Size**	**#Attribute**	**#Classes**	**(minority/ majority)**
Haberman	306	3	2	81/225
German	1000	24	2	300/700
Pima	768	8	2	268/500

2.2 Benchmark for Imbalanced Data Sets

In order to validate the proposed technique, several benchmark data sets from the UCI database repository [20] are used as shown in Table 2. Only binary classification data set is used in this study and it must have the imbalanced ratio between the two classes.

3. Methodologies

To begin with, the dataset is divided randomly in portion of 60% as training samples and 40% as testing samples. The configuration of a single layer feedforward neural network is shown in Fig. 1. The setup configuration is based on the trial and error procedure to find a good number of neuron, *m,* in hidden layer (HL). Levenberg-Marquardt (LM) backpropagation algorithm is chosen as training algorithm to minimize mean square error (MSE) between the actual outputs of the network and the desired outputs. The LM is chosen as the training algorithm due to the higher prediction as reported by Giang [1] and the fastest algorithm in the MATLAB's toolbox than other algorithms. Giang *et al* has done an investigation with selected benchmark data sets form UCI machine learning repository using different training algorithm and his report indicates that LM gives a higher prediction performance than others for imbalanced data sets. By using neural network, the training samples are trained through the designed network and the all available samples are tested. By referring in Fig. 2, the desired predicted output that comes from the trained network is used as an additional input (domain expert) to the naïve bayes classifier.

Naïve Bayes classifier was initially introduced by Duda and Hart in 1973. This illustrates a classical classifier in Bayesian Network and so-called "naïve" Bayesian classification [21] due to the inputs parameters are independent between each other given by the class, S. The Naïve Bayes network is fixing between inputs, Yn= (Y1, Y2, Y3 . . . Yn) and output, S as shown in Fig.3.

At this stage, continuous-value input parameter in Eq. (4) is chosen to generalize the classification system as learning mechanism to all available imbalanced data sets.

$$f(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \quad (4)$$

Where, x is a value of sample in each row. μ is a mean of input samples and σ^2 is a variance of input sample. Then, the probability of P(Y) given by a class, S=s in the following equation,

$$P(Y) = p(S = s) \prod_{i=1}^{n} p(Y_i = y_i) | S = s) \quad (5)$$

The prediction process is done by using a maximum argument in Eq. (6), after the probability of each input P (Yi), Eq. (5) is calculated.

$$Y^{predict} = argmax_s P(Y) \quad (6)$$

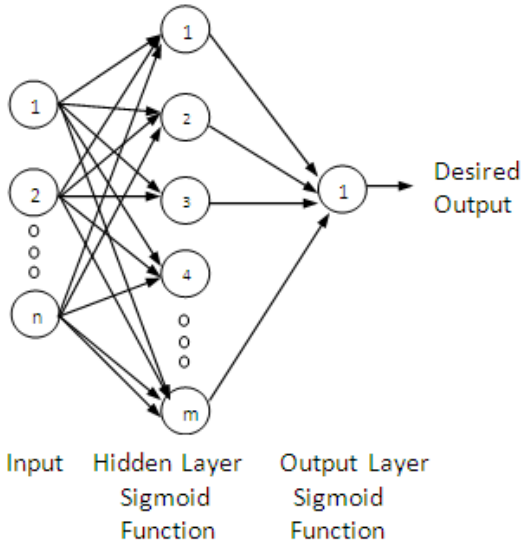

Fig 1: A single layer feedforward neural network

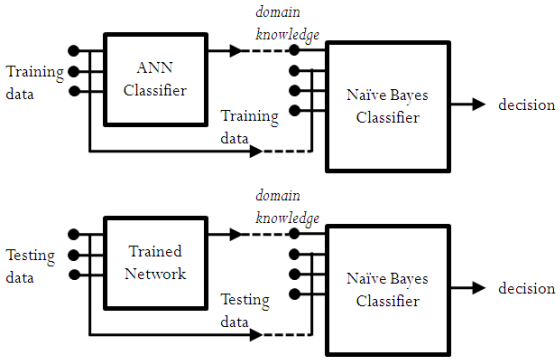

Fig 2: Integration of expert's knowledge in Naïve bayes classifier.

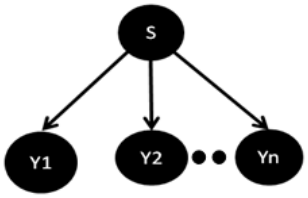

Fig 3: The Structure of Naïve Bayes Network

4. Experimental Results and Discussions

The experiments were performed using the Pentium (R) Dual-Core 2.60GHz computer, with 2GB of RAM. There are three selected benchmark data sets used as mention in Section 4, Haberman Survival, German Credit and Pima Indian data sets. The experiment compares a prediction performance (Gmean) between the conventional naïve bayes classifier and the decision support system designed for imbalanced dataset as shown in Fig. 2. The type of knowledge for all data sets is a data-driven guided by expert system (neural network) as discussed in Section 5. In general, the presented results in Table 3 are improved by 10 percent by incorporating additional knowledge (domain knowledge) to the naïve bayes classifier which is driven by neural network. The accuracy of the Haberman Survival data set is improved by using the proposed technique from 53% to 61.2% while German Credit dataset is improved from 65% to 71.09 of the performance accuracy. For Pima Indian data set, 10% accuracy improvement is recorded as compared to the naïve bayes classifier.

Based on these results, the improvement of accuracy is illustrated without modification at data level and algorithmic level. A set of degree improvement is required in internal algorithm that can solidify the model with the addition of any pre-processing technique for the ignorant imbalanced data set.

5. Conclusion and Future Work

This study briefly introduces a way to incorporate expert knowledge to the expert system based on the data-driven technique. The combination of conventional and classical classifier is able to guide a main classifier to achieve better prediction performance. This study also realizes that the knowledge engineering and machine learning is in different discipline but the combination of both discipline make a system more informative. Data-driven type reduces the computational time and the designing process, while using expertise-driven types, it takes a long time particularly to find relevant rules in learning process for classification. In future, making a Bayesian network (BN) is more applicable for imbalanced datasets, the network of BN will be defined expert's knowledge.

Acknowledgement

This research was supported by the Grant from UTM-INTEL Research Collaboration, Vot 73332.

Table 3. Comparison results

Data Sets	Type of Knowledge	Bayesian Network (Naïve Bayes)	
		Conventional Naïve Bayes Classifier (%)	Incorporate additional knowledge (%)
Haberman	Machine (ANN)	53	61.2
German	Machine (ANN)	65	71.09
Pima	Machine (ANN)	65	75

References

[1] Giang H. Nguyen, Abdesselam Bouzerdoum, and Son L. Phung, "A Supervised Learning Approach for Imbalanced Data Sets", Proceeding 19th International Conference on Pattern Recognition (ICPR 2008), (2008), pp. 1-4.

[2] Friedman, N., Geiger, D. and Goldszmidt, M., "Bayesian Network Classifiers", Machine Learning, Vol. 29, (1997), pp. 131-161.

[3] VicencSoler, Jesus Cerquides, JosepSabria, JordiRoig and Marta Prim, "Imbalanced Datasets Classification by Fuzzy Rule Extraction and Genetic Algorithms", Proceeding in Sixth IEEE International Conference on Data Mining-Workshop (ICDMW' 06), (2006), pp. 330-336.

[4] Kaizhu Huang, Haiqin Yang, Irwin King, Michael R. Lyu, "Learning Classifiers from Imbalanced Data Based on Biased Minimax Probability Machine", IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR04), (2004).

[5] Zhou Quan, Gu Lin-Gang, Wang Chong-Jun, Wang-Jun and Chen Shi-Fu, "Using An Improved C4.5 for Imbalanced Dataset of Intrusion", Proceedings of the 2006 International Conference on Privacy, Security and Trust, Vol. 380, (2006), pp. 1-4.

[6] Yuchun Tang, Yan-Qing Zhang, Nitesh V. Chawla, and Sven Krasser, "SVMs Modeling for Highly Imbalanced Classification", IEEE Transaction on System, Man, and Cybernetics, Vol. 39, Issue 1, (2008), pp. 281-288.

[7] D. Cieslak, N. Chawla, and A. Striegel, "Combating Imbalance in Network Intrusion Datasets", Proceeding of IEEE International Conference on Granular Computing, (2006), pp. 732-737.

[8] B. Anuradha and V. C. Veera Reddy, "ANN for Clasification of Cardiac Arrhythmias", Asian Research Publishing Network (ARPN) Journal of Engineering and Applied Sciences, Vol. 3, No. 3, (2008), pp. 1-6.

[9] L. Bruzzone, S.B. Serpico, "A Classification of imbalanced remote-sensing data by neural networks", Pattern Recognition Letters, Vol. 18, (1997), pp. 1323-1328.

[10] Yi Lu, Hong Guo, and Lee Feldkamp, "Robust Neural Learning from Unbalanced Data Samples", Proceedings of IEEE International Joint Conference on Neural Networks, IEEE World Congress on Computational Intelligent, Vol. 3, (1998), pp. 1816-1821.

[11] WK Yip, KG Law, and WJ Lee, "Forecasting Final/Class Yield Based on Fabrication Process E-Test and Sort Data", Proceedings of IEEE International Conference on Automation Science and Engineering, (2007), pp. 478-483, Scottsdale AZ.

[12] N.V.Chawla, L.O.Hall, K.W.Bowyer, and W.P.Kegelmeyer, "SMOTE: Synthetic Minority Oversampling Technique", Journal of Artificial Intelligence Research, Vol.16, (2002), pp. 321-357.

[13] H.Han, W.Y.Wang, and B.H.Mao, "Borderline- Data Sets Learning", in Proceedings of the International Conference on Intelligent Computing 2005, Part 1, LCNS 3644, (2005), pp 878-887.

[14] R. Dybowski, K.B. Laskey, J.W. Myers and S. Parsons, "Introduction to the special issue on the fusion of domain knowledge with data for decision support", Journal of Machine Learning Research 4, Vol. 4, (2003), pp. 293-294.

[15] Atish P. Sincha and H. Zhao, "Incorporating domain knowledge into data mining classifier: An application in Indirect lending". Elsevier, Decision Support System 46, (July 2008), pp. 287-299.

[16] S.S. Anand, D.A. Bell, J.G. Hughes, "The Role of Domain Knowledge in Data Mining", Proc. 4th Int'l. ACM Conf. on Information and Knowledge Management, (1995), pp. 37-43.

[17] Yi L. Murphey, Haoxing Wang, GuobinOu, Lee A. Feldkamp, "OAHO: an Effective Algorithm for Multi Class Learning from Imbalanced Data", IEEE, International Joint Conference on Neural Networks, Orlando, Florida, USA, 2007.

[18] Z.Q. Zhao, A novel modular neural network for imbalanced classification problems, Pattern Recognition Letters, Vol. 30, (2008), pp. 783-788.

[19] Cheng G. Weng and Josiah Poon, "A New Evaluation Measure for Imbalanced Datasets", (2006). Copyright c 2008, Australian Computer Society, Inc. This paper appeared at conference Seventh Australasian Data Mining Conference (AusDM 2008), Glenelg, Australia. Conferences in Research and Practice in Information Technology, Vol. 87. John F. Roddick, Jiuyong Li, Peter Christen and Paul Kennedy, Eds. Reproduction for academic, not-for profit purposed

permitted provided this test is included.
[20] UCI Machine Learning Repository [http://www.ics.uci.edu/~ learn/MLRepository.html]. Irvine, CA: University of California, School of Information and Computer Science.
[21] Duda, R. O. and P. E. Hart, "Pattern Classification and Scene Analysis", New York, John Wiley & Sons. 1973.