Proceedings of the 1^{st} KMITL International Conference on Integration of Science and Technology for Sustainable Development, August 25-26, 2004, Lad Krabang, Bangkok, Thailand, pages 275-278

The Impact of Noise at Different Data Attributes

Nittaya Kerdprasop, Kittisak Kerdprasop, Laksamee Khomnotai and Thammasak Thianniwet

Data Engineering and Knowledge Discovery (DEKD) Research Unit
School of Computer Engineering
Suranaree University of Technology
Nakhon Ratchasima 30000, Thailand
E-mail: nittaya, kerdpras@ccs.sut.ac.th

ABSTRACT

Real-world data often suffer from corruptions or noise. The most serious negative impact of noise is that it can reduce machine learning performance in terms of learning accuracy. Most learning algorithms have integrated various approaches to handle noisy data. However, rare research has been conducted to systematically explore the impact of noise, especially when noise occurs at different attributes. We investigate the effect of class noise, noise in principal attributes, and noise in irrelevant attributes to the learning accuracy. Our conclusions can be served as a preliminary step toward the designing of handling mechanisms for a specific kind of noise.

Keywords

Noise, Class noise, Attribute noise, Noise impact

1. INTRODUCTION

Noise is a random error in data. Noisy data contain incorrect attribute values caused by many possible reasons, for instance, faulty data collection instruments, human errors at data entry, errors in data transmission. If noise occurs in the training data, it can lower the performance of the learning algorithm. The most serious effect of noise is that it can confuse the learning algorithms to produce complex and distorted results. The long and complex results are dued to the attempt to fit every training data into the concept descriptions. This situation is named the overfitting problem.

Most learning algorithms are designed with the awareness of noisy data. Thus, there exist some mechanisms in dealing with noise, for example, the ID3 algorithm[3] uses the prepruning technique to avoid growing a decision tree too deep down to cover the noisy training data. Some algorithms adopt the technique of postprocessing to reduce the complexity of the learning results. Postprocessing technique includes the cost-complexity pruning, reduced error pruning, and pessimistic pruning described in Quinlan[4,5].

Even though most existing learning algorithms include various noise-handling techniques, the existence of noise can still affect the learning results negatively. The focus of this paper is to observe the impact of noise to the learning algorithms. We categorize noise into three groups: class noise, noise in principal attributes, and noise in irrelevant attributes. We investigate the relationship between various groups of noise and learning accuracy. Our conclusions can be used to enhance the handling techniques specific to the noise of different types.

2. EXPERIMENTAL METHODOLOGY

We study the impact of noise on three data sets: Monk1 (124 instances), Ionosphere (234 instances), and Chess (2,130 instances). These data sets are UCI[2] standard data for testing the performance of machine learning algorithms. We generate noise varying from 0% to 45% on different groups of attributes. Class noise[1] is an occurrence of random error in target attribute (i.e., the instances are mislabeled). Attribute noise is a random error in predicting attributes. Every attribute except a class attribute has an equal probability of noise occurrence. Instead of simply studying attribute noise, we extend our investigation to the impact of noise that occurs in the principal attributes, i.e., attributes highly correlate to class prediction, and the impact of noise if it occurs at less-relevant attributes.

We test the impact of noise on two learning algorithms: naive Bayes, and neural network. These algorithms are known as a noise-tolerant system. We compare their noise-tolerant performance when noise is introduced to the class attribute, the principal predicting attributes, and attributes irrelevant to the prediction. For consistency, we supply the same data set to each test.

3. RESULTS AND DISCUSSIONS

The tolerance against class noise of naive Bayes and neural network algorithms tested on three data sets is shown in Figure 1. The effects of noise in principal attributes and irrelevant attributes are shown in Figures 2 and 3, respectively.

The experimental results reveal that:

(1) Class noise has more impact on the neural network algorithm than on the naive Bayes. This negative effect can be noticed clearly on a large data set (Chess data).

(2) When noise occurs in highly predictive (or principal) attributes, it has much more effect on the neural network algorithm than the naive Bayes.

(3) For the small data set (Monk1), noise in irrelevant attributes has no effect on both algorithms. But on larger data set, this kind of noise can slightly degrade the performance of the neural network algorithm.

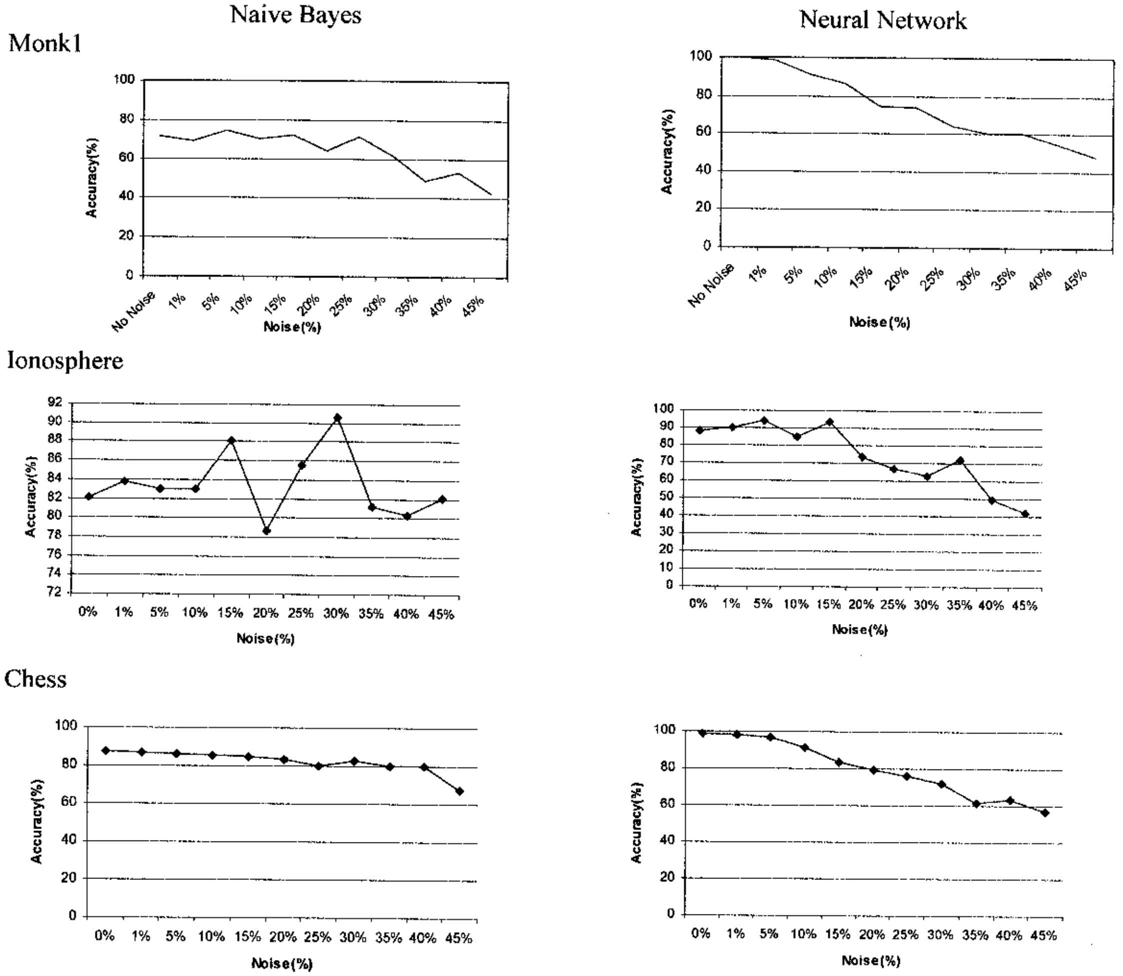

Figure 1. The effect of class noise

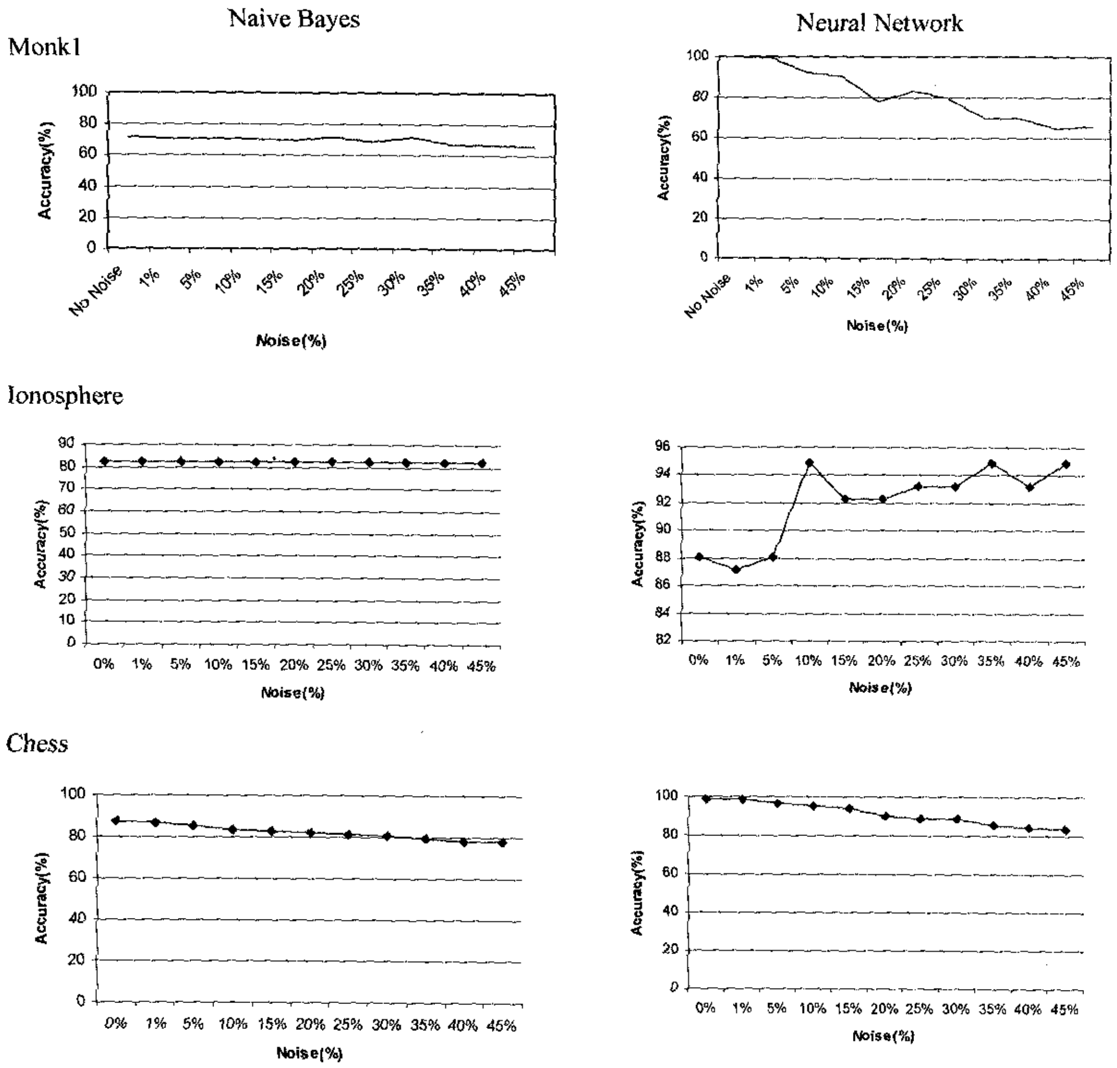

Figure 2. The effect of principal attribute noise

4. CONCLUSION

Noise in a data set can happen in different forms: (1) misclassification or wrong labeled instances, (2) errorneous or distorted attribute values, (3) contradictory or duplicate instances having different labels, (4) missing attribute values. All kinds of noise can more or less affect the learning performance. We specific our investigation to the fist two kinds of noise, which are termed class noise and attribute noise, respectively. Class noise has been studied extensively by many researchers, whereas attribute noise is less thoroughly studied. We extend the study on attribute noise by categorize it further to principal attribute noise and irrelevant attribute noise. We find that principal attribute noise has more negative impact on the learning performance than the class noise. Noise in irrelevant attributes can somehow affect the neural network algorithm. Our future research is to extend our study and to design a handling mechanism specific to each kind of noise.

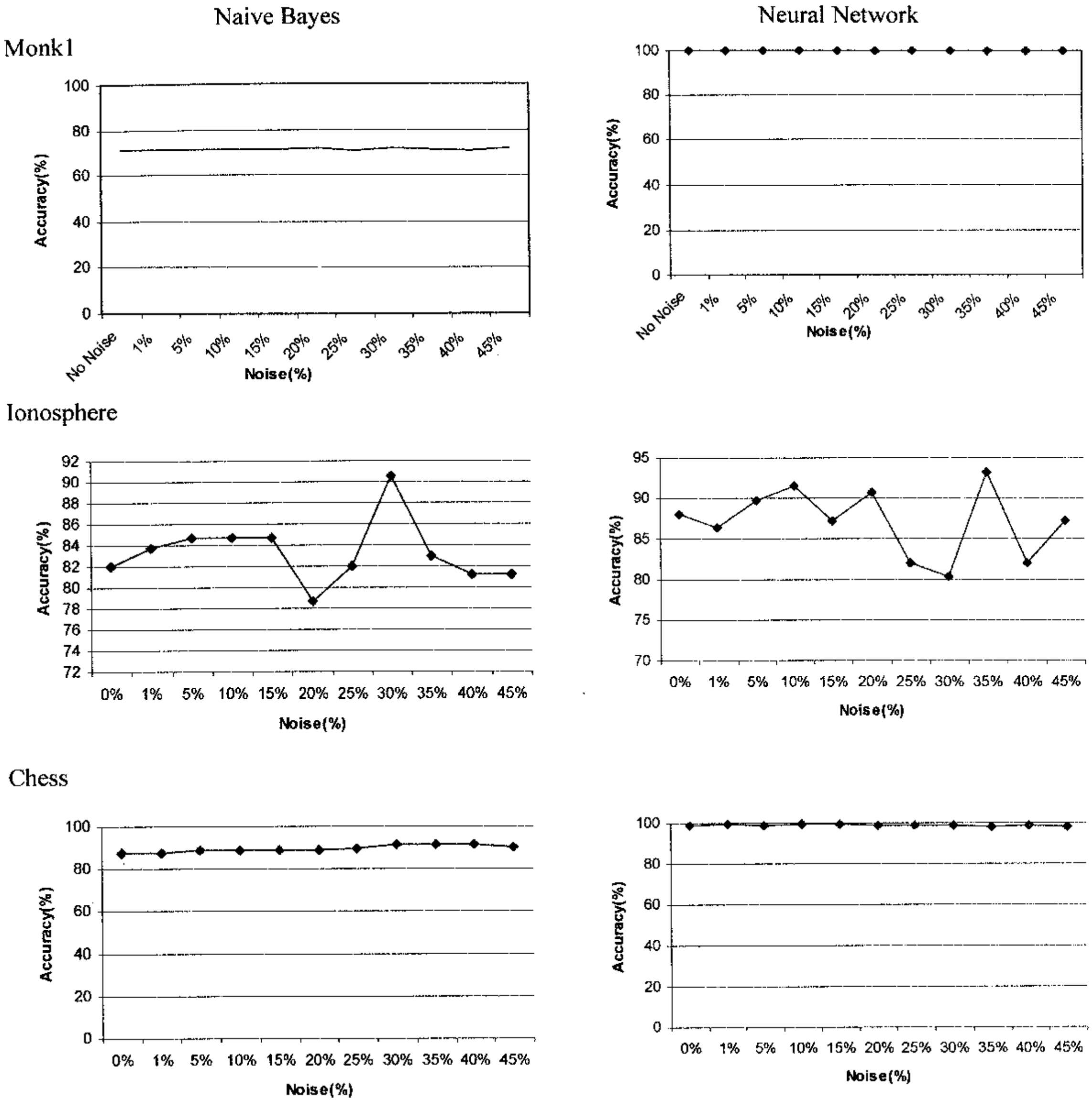

Figure 3. The effect of irrelevant attribute noise

5. REFERENCES

1. Angluin, D., and Laird, P. (1988). *Learning from noisy examples*. Machine Learning, 2, 343-370.

2. Merz, C.J., and Murphy, P.M. (1997). *UCI Repository of machine learning database*. http://www.ics.uci.edu/~mlearn/MLRepository.html

3. Quinlan, J.R. (1986). *Induction of decision tree*. Machine Learning, 1, 81-106.

4. Quinlan, J.R. (1989). *Simplifying decision tree*. In B. Gaines and J. Boose (Editors), Knowledge Acquisition for Knowledge Based Systems, Vol. 1, Academic Press.

5. Quinlan, J.R. (1992). *C4.5: Programs for Machine Learning*. Morgan Kaufmann.