SENSITIVITY OF DIFFERENT MACHINE LEARNING ALGORITHMS TO NOISE[*]

Abhinav Atla
Computer Science Department
University of Central Arkansas
Conway, AR 72034
abhinav655@gmail.com

Rahul Tada
Computer Science Department
University of Central Arkansas
Conway, AR 72034
t.rahulreddy@gmail.com

Victor Sheng
Computer Science Department
University of Central Arkansas
Conway, AR 72034
ssheng@uca.edu

Naveen Singireddy
Computer Science Department
University of Central Arkansas
Conway, AR 72034
naveen.singireddy@gmail.com

ABSTRACT

Noise in data is an effective cause of concern for many machine learning techniques that are used in modeling data. Researchers have studied the impact of noise only on some particular learning algorithm, but only very few attempted to analyze the effect of noise on different ones. In this work, we study the noise sensitivity of four different learning algorithms under different intensities of noise. Particularly, we compare the noise sensitivity of decision tree, naïve bayes, support vector machine, and logistic regression. The algorithms are tested on different datasets that are artificially injected with different degrees of noise. The study helps us understand the impact of different levels of noise on the learning algorithms mentioned above. Furthermore, it also guides of choosing the learning algorithms. In general, naïve bayes is the most resistant to noise. However, it performs also the worst. The other algorithms perform much better than naïve bayes especially after the noisy level is lower than 40%. When we have approaches to improve the data quality (reduce the noise level), decision tree is the most preferred one, followed by support vector machine and logistic regression, not naïve bayes.

1. INTRODUCTION

Data extracted from real-world problems using supervised learning techniques frequently contains noise. Noise decreases the quality of the data and might affect the learning process leading to inaccurate models. However, the supervised learning algorithms are developed based on the assumption of clean data. This research is to discover the performance of learning algorithms with noisy training data. It will investigate the noise sensitivity of learning algorithms with various percentages of noise. The results of this research can assist choosing proper learning algorithms. It can recommend which learning algorithm is preferred under what level of noise. Furthermore, if we can reduce the noise level, which algorithm should be chosen? In this work we are concerned with inaccuracies that migrate from data contaminated with random errors. Random errors can be defined as the uncertainty detected by executing learning algorithms with different conditions of noise. In this work we utilize classifiers such as decision tree, naïve bayes, support vector machine, and logistic regression to perform classification efficiently on different datasets. We study the performance of these algorithms under different noise levels from 0% (no noise) to 50% (very noisy).

The paper is organized as follows. A brief summary of the related literature on handling noise is provided in section 2. Section 3 reviews the different learning algorithms used to analyze their sensitivity under different noise levels. Sections 4 and 5 presents an explanation of the mechanism used for generating noise and the analysis of the experimental results for the four machine learning algorithms. Finally Section 6 provides conclusions about the performance of the algorithms on different datasets and various levels of noise.

2. RELATED WORK

Real-world data always contain noise. In order to apply learning algorithms, data preprocessing is the first step. There are many tasks during data preprocessing. Among them, data cleaning is one of the important tasks. There are many works on detecting noise and reducing noise (Xiong 2006; Zhu et al., 2003, 2006; Kubica 2003; Brodley 96, 99). There are two types of noise: attribute noise and class noise (Zhu et al., 2003, 2006). Brodley (Brodley 96, 99) detected and removed the class noise. Most of the work has been done on noise detection focused on supervised classification problems. However, Xiong et al. (Xiong et al., 2006) studied the approach of remove noise for unsupervised learning. Kubica & Moore (Kubica 2003) studied how to identify and remove the attribute noise, such that the remaining information in the training examples can still be used in modeling. Different from previous work, this paper focuses on the impact of class noise on supervised learning algorithms and investigates the noise sensitivity of different learning principles through analyzing the performance of learning algorithms under different level of noise. Thus, we can choose the proper learning algorithm. This is also needed even if the detecting and reducing noise approaches are applied to improve the data quality.

3. LEARNING ALGORITHMS

There exist more than 20 different learning algorithms, including the basic ones and improved variations. According to the categorization of WEKA (Witten and Frank), we choose the fundamental one from each category. That is, we are going to investigate the noise sensitivity of four basic learning algorithms. They are decision tree, naïve bayes, support vector machine, and logistic regression.

A decision tree algorithm (DT in short) partitions the input space into small sets, and labels them with one of the various output categories. That is, it iteratively selects a specific attribute to extend the learning model. According to the values of the specific attribute, a large group of cases are categorized into sub-groups. The essence of the algorithm is the criteria of selecting a specific attribute from the set of attributes available. There exist several criteria, such as accuracy improvement, entropy reduction, information gain, and gain ratio (details of these criteria can be found in (Mitchell machine learning book)). The noisy label information will directly affect the estimation of all the criteria. Thus, the model built on decision tree algorithm would be affected by the noisy label.

Naïve bayes (NB in short) (John and Langley 1995) is based on bayes theorem. Specifically, it is based on the properties of estimating the frequency of each value of every attribute under a given class from the obtained dataset. We can image that there is no difference of these estimation with/without noisy labels if both the class distribution and the distribution of the noisy label follow the complete random distribution. That is, naïve bayes is noisy-insensitive, particularly in estimating the conditional probabilities of each attribute value under given classes. We will see that there is little difference on the performance of naïve bayes under different noisy levels. Of course, it is not always true that the conditions are satisfied on the real world datasets. Noise still impacts the performance of the model built on naïve bayes.

Support Vector Machine (SVM in short) is one of the kernel approaches for classification. It constructs a hyperplane in high dimension space to classify cases into different classes. Its intuition is to find the hyperplane that has the largest distance to the nearest training cases. These nearest training cases are commonly referred as support vectors. According to the vectors on each side, a sub-hyperplane can be built for each side. The maximum margin between the two sub-hyperpanes is achieved to reduce the general classification errors. When the data have noise, it is possible that these support vectors could have noise too. Thus, the hyperplane and the two sub-hyperplanes (found based on support vectors) can vary. That is why support vector machine is very sensitive to noise. We further discuss this after describing experimental results.

Logistic regression (logistic in short), like naïve bayes discussed above, is part of a category of statistical models called generalized linear models. However, unlike the independent assumption among the variables in naïve bayes, logistic regression has no such assumption. In addition, it also makes no assumption about the distribution of the independent variables. Although both logistic regression and naïve bayes are statistical models, the basic ideas of them are different. Logistic regression estimates the probability of being positive case (for binary classification). It can be inferred that the probability migrates toward the uniform distribution when more noise labels are involved. Thus, it is more noise sensitive. The experimental results also show this (refer to Section 5).

4. EXPERIMENTAL SETUP

To study the noise sensitivity of the four different learning algorithms (decision tree, naïve bayes, support vector machine, and logistic regression), we run experiments on all the classification datasets downloaded from WEKA website. In the experiments, we assume that these datasets are clean (no noise, especially no noise class labels). In order to have the datasets to study the noise sensitivity, we inject noise into these datasets. Since we focus on investigation of the class noise, we only introduce noise into the class labels.

Here is the simulation procedure:

```
1. Repeat 10 times
    a. Divide a dataset into training part (70%) and test part (30%)
    b. Keep the original class labels in array for all training examples in the
       training part
    c. For each training example (simulate the noise labels with the control of
       the noise level)
        i. If its label is wrong, randomly generate a label from the labels other
           than the original label to replace the original label
    d. Build classification model using a specific learning algorithm
    e. Make prediction for the test examples in test part, and output the
       classification accuracy
2. Output the average classification accuracy over 10 repeats
```

We studied the noise sensitivity on the four learning algorithms step by step. We first focus on the impact of noise labels on binary classification. The binary datasets from WEKA website are *bmg, expedia, kr-vs-kp, mushroom, qvc, sick, spambase, tic-tac-toe*, and *travelocity*. We further study the performance of each learning algorithm for multiclass classification to see whether the noise impact on the performance on binary classification stays. The multiclass datasets from WEKA website are *anneal, audiology, autos, balance, glass, heart-c, heart-h, iris, letter, lymph, primary-tumor, segment, splice, thyroid, vehicle, vowel, waveform*, and *zoo*. Because of the page limitation, we could not show the properties of these datasets here.

From the procedure above, we can see that we have a parameter (noise level) to control the amount of noise introduced. It should be noted that the noise introduced here are random. In details, in the step 1.c.i of the simulation procedure above, for each training example, we randomly generate the noise labels to replace the original labels in the training data. Specifically, for binary classification, if the label of a training example is wrong, its opposite label should be used to replace the original label. For multiclass classification, any one from the rest labels (different from the original label) has the same probability to be used to replace the original one.

5. EXPERIMENTAL RESULTS

In this section, we compare the performance of different machine learning algorithms on the datasets with different percentages of label noise injected. Noise degree is controlled from 50% to 10%. In addition, to see the noise label impact, we also have the experimental results for each dataset under the four learning algorithms without noise. Evaluation of the machine learning algorithms performance is done by comparing the average test accuracy of the models constructed on the training dataset and tested on the test dataset. The results on the binary datasets are shown in Table 1. The average results overall all nine binary dataset are shown in Figure 1. And the results of the multiclass datasets are shown in Table 2.

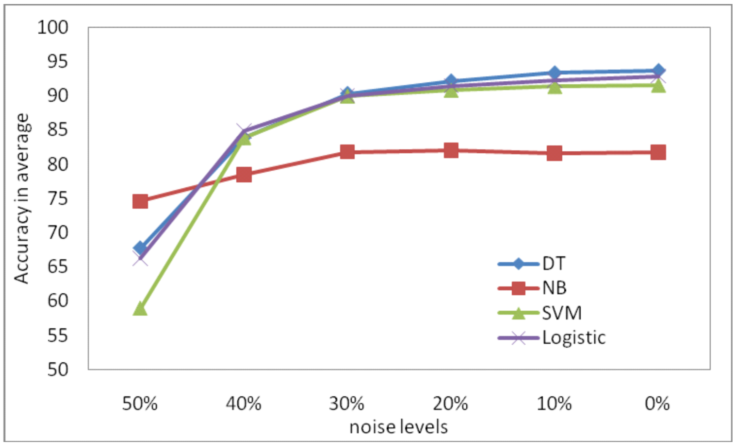

Figure 1: The average accuracy over the nine binary datasets for the four learning algorithms.

From Figure 1 and Table 1, we can conclude that naïve bayes (NB) is the most noise tolerant among the four learning algorithms. Its performance does not change much when the noise level decreases (that is, the label quality increases). Its curve over the noise level is almost flat. It is obvious that naïve bayes always performs the best in average when the noise level is the highest (50%), followed by decision tree (DT) and logistic regression (logistic), followed by support vector machine (SVM). However, when the noise level reduces, other three algorithms perform better. Especially, when the noise level reduces from 50% to 40%, the performance of all the three algorithms improves quickly. Although their performance keeps improving when the noise level continues to reduce, the acceleration of the improvement slows down. Among the three noise sensitive learning algorithms, decision tree performs the best, followed by logistic regression,

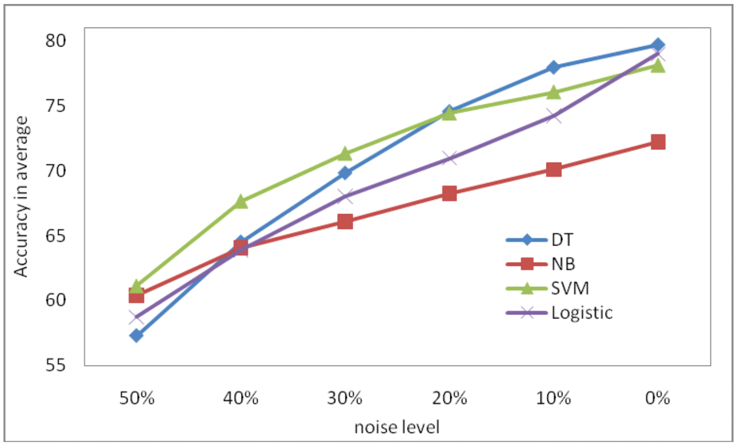

Figure 2. The average accuracy over the 18 multiclass datasets for the four learning algorithms.

followed by support vector machine. In other words, if we have approaches which can improve the quality of the labels, decision tree is the first preference, followed by logistic regression and support vector machine.

Dataset	%Noise	50%	40%	30%	20%	10%	0%
bmg	DT	*80.23*	*81.70*	*82.94*	*85.20*	*86.28*	*86.07*
	NB	69.92	67.60	68.92	66.74	66.41	65.54
	SVM	76.87	76.87	76.87	76.87	76.87	76.94
	Logistic	76.87	76.81	77.59	78.25	79.08	78.95
expedia	DT	87.39	87.42	90.61	*91.97*	*92.89*	*93.34*
	NB	85.88	85.68	85.93	85.40	84.60	83.54
	SVM	87.45	87.77	88.59	90.85	91.42	91.88
	Logistic	*89.20*	*90.04*	*91.11*	91.32	91.55	91.55
Kr-vs-kp	DT	*72.55*	*83.86*	*95.70*	*98.22*	*99.12*	*99.24*
	NB	66.48	72.93	79.58	83.88	86.10	87.30
	SVM	57.55	83.26	90.65	90.86	94.16	95.38
	Logistic	60.99	73.36	85.71	92.11	95.22	97.44
mushroom	DT	74.69	90.23	99.61	99.87	*100.00*	*100.00*
	NB	*96.14*	96.71	97.32	97.56	97.21	95.87
	SVM	74.09	*97.89*	*99.86*	*99.98*	99.98	*100.00*
	Logistic	74.63	96.07	99.74	99.93	99.98	*100.00*
qvc	DT	*83.97*	*85.47*	*86.95*	*87.81*	*88.88*	*88.95*
	NB	69.24	69.77	68.03	67.02	65.86	65.92
	SVM	82.40	82.40	82.54	82.84	83.27	83.26
	Logistic	82.79	82.98	83.50	83.64	83.26	83.12
sick	DT	*47.52*	*84.08*	*95.03*	*97.24*	*98.62*	*98.75*
	NB	40.37	64.17	83.16	82.93	88.74	92.14
	SVM	24.79	80.52	93.58	94.08	94.08	94.08
	Logistic	35.59	84.04	94.18	95.20	95.28	96.61
spambase	DT	47.57	*86.65*	*88.09*	89.41	91.17	92.38
	NB	*85.22*	86.02	86.36	85.35	77.22	79.45
	SVM	41.68	59.72	87.49	90.43	90.58	90.68
	Logistic	55.77	76.96	86.40	*90.46*	*91.72*	*92.88*
tic-tac-toe	DT	65.51	71.74	77.77	80.24	83.07	84.18
	NB	67.32	69.16	71.71	73.94	73.07	70.94
	SVM	65.75	*95.61*	*96.48*	*97.70*	*98.22*	*98.22*
	Logistic	*80.59*	91.15	95.26	96.17	97.91	*97.91*
travelocity	DT	50.09	83.14	*95.87*	*99.08*	*99.63*	*99.71*
	NB	*90.54*	*94.32*	94.90	94.87	94.89	95.01
	SVM	19.64	90.47	93.02	93.09	93.10	93.18
	Logistic	40.02	91.89	96.45	95.81	95.76	97.20
average	**DT**	**67.72**	**83.81**	**90.29**	**92.12**	**93.30**	**93.62**
	NB	**74.57**	**78.48**	**81.77**	**81.97**	**81.57**	**81.75**
	SVM	**58.91**	**83.83**	**89.90**	**90.74**	**91.30**	**91.51**
	Logistic	**66.27**	**84.81**	**89.99**	**91.43**	**92.20**	**92.85**

Table 1. The accuracy of the four learning algorithm on binary classification.

We further study the noise sensitivity of the four learning algorithms with multiclass classification. We have done the experiments on 18 datasets. Because of the space limitation, we could not show their results here. We just show the overall average in term of accuracy of the 18 datasets in Figure 2. Figure 2 verifies the conclusions we made from the experimental results for binary datasets: naïve bayes is the most noise tolerant learning algorithm and it performs well when the noise level is the highest (50%). Although it performs better when the noise level reduces, the performance of the other three learning algorithms improves significantly. When the noise level is higher than 20%, support vector machine performs the best in average, followed by decision tree, followed by logistic regression. However, when the noise level is further reduced, decision tree performs the best, followed by support vector machine, followed by logistic regression.

6. CONCLUSION

In this research, we studied on how the quality of models is affected by different amounts of noise for different machine learning algorithms. The study was performed on four different classifiers called decision tree, naïve bayes, support vector machine, and logistic regression. A detailed experimentation proves that the behavior or each algorithm depends on the percentage of noise injected and the characteristics of different datasets.

This kind of a study is very useful in situations of real world data processing that may contain implicit and explicit errors. The results show that naïve bayes is the most noise tolerant algorithm. However, decision tree performs the best in average under different noise level for most datasets (binary and multiclass), followed by logistic regression and support vector machine. When we have approaches to improve the data quality (reduce the noise level), they are more preferred than naïve bayes.

7. REFERENCES

[1] Brodley, C.E. and Friedl, M.A. Identifying and eliminating mislabeled training instances, *Proc. of 13th National Conf. on Artificial Intelligence*, 1996, pp.799-805.

[2] Brodley, C.E. and Friedl, M.A. Identifying mislabeled training data, *Journal of Artificial Intelligence Research*, 1999, 11, 131-167.

[3] John, G. H., and Langley, P. Estimating Continuous Distributions in Bayesian Classifiers. In *Eleventh Conference on Uncertainty in Artificial Intelligence*, San Mateo, 338-345, 1995.

[4] Kubica, J., and Moore, A. Probabilistic Noise Identification and Data Cleaning, pp.131-138, *Proceedings of the third IEEE International Conference on Data Mining (ICDM'03)*, 2003

[5] Witten, I.H., and Frank, E. *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd ed. Morgan Kaufmann Publishing, 2005.

[6] Xiong, H., Pandey, G., Steinbach, M. and Kumar, V. Enhancing data analysis with noise removal. *IEEE Transactions on Knowledge and Data Engineering*, 2006, 18, 304-319.

[7] Zhu, X., Wu, X. and Chen, Q. Eliminating Class Noise in Large Datasets. *Proceedings of the 20th ICML International Conference on Machine Learning (ICML 2003)*. Washington D.C., 2003, pp. 920-927.

[8] Zhu, X., Wu, X. and Chen, Q. Bridging Local and Global Data Cleansing: Identifying Class Noise in Large, Distributed Data Datasets. *Data Mining and Knowledge Discovery*, 2006, 12, 275-308.