An Empirical Study of Class Noise Impacts on Supervised Learning Algorithms and Measures

Victor S. Sheng, Rahul Tada, and Abhinav Atla
Department of Computer Science, University of Central Arkansas, Conway, AR, USA

Abstract - *Noise in data is an effective cause of concern for many machine learning techniques. Researchers have studied the noise impacts only on some particular learning algorithm. We empirically study the noise impacts on four different representative learning algorithms and the two popular measures (accuracy and AUC) under different intensities of noise, particularly decision tree, naïve bayes, support vector machine, and logistic regression. Our empirical results show that AUC is more tolerant to noise. Among the four algorithms, naïve bayes is the most resistant to noise, but it performs the worst in accuracy. The other algorithms perform much better than naïve bayes especially after the noisy level is lower than 40%. When we develop approaches to improve the data quality (reduce the noise level) and build model with higher accuracy, decision tree is the most preferred one, followed by logistic regression and support vector machine. However, logistic regression performs the best in AUC.*

Keywords: Noisy learning, noise impact, supervised learning, machine learning, data mining

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. This research is to discover the performance of learning algorithms with noisy training data, and 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 could bring higher benefit?

Cleaning data and improving the data quality is one of the preprocessing of KDD process [1]. Cleaning the mislabeled instances before training a learning model is very common. Mislabeled instances hurt the performance of supervised learning algorithms. Great efforts contribute to identify the potential mislabeled data and further to handle these mislabeled data properly [2] [3]. There are several approaches addressed to improve the label quality of training instances. The knowledge of the impacts of noise on different supervised learning algorithms can help us choose the learning algorithms, which are sensitive to noise reduction and perform better with the data quality improvement. On the other hand, try not to choose the noise insensitive algorithms.

Real-world data contains attribute noise and/or class noise. We focus on the class noise. Specifically, we are concerned with inaccuracies that migrate from data contaminated with class noise. We study the performance of four representative algorithms: decision tree, naïve bayes, support vector machine, and logistic regression, under different noise levels from 0% (no noise) to 50% (complete noise for binary classification). This empirical study guides the algorithm choice under different noise levels, particularly the two extreme cases: very noise or no noise at all. Our empirical results show that naïve bayes is the most resistant to noise, but it performs the worst in accuracy. The other algorithms perform much better than naïve bayes especially after the noisy level is lower than 40%. Decision tree is the preferable one. However, in AUC (area under the ROC curve) [4] [5] [6], naïve bayes performs not bad, although it follows logistic regression. Logistic regression is the best.

The investigation of the noise sensitivity of the basic algorithms under different noise levels is the preliminary study, which helps the study of improving data quality (reduce the noise level via removing/correcting identified noise). Studying data quality improvement needs to choose noise reduction sensitive algorithms. There exist numerous articles about handling noisy data, including removing the noise instances directly and correcting the noise instances. But, according to our knowledge, none of them emphasizes the difference of noise impacts on different learning algorithms, and none of them explicitly explained why the certain algorithms were chosen. Either for removing or for correcting, it can be expected that the noise reduction sensitive learning algorithms would be proper ones to be chosen, not the noise reduction insensitive learning algorithms. Our experimental results show that decision tree is the most preferred one, followed by logistic regression and support vector machine, in accuracy.

This paper also investigates the noise sensitivity of the measures under different noise levels. Accuracy and AUC are two common measures for supervised learning algorithms. In this paper, we investigate their noise sensitivity via comparing the performance of the supervised learning algorithms based on these two measures. It has been shown that the two measures are not completely identical [7]. Our experimental results show that accuracy is more sensitive to noise, comparing to AUC. That is, AUC is more tolerant to the noise impact on learning algorithms, particularly on naïve bayes and logistic regression. When we study the data improvement, accuracy is more preferable than AUC.

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

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 exists much work on data cleaning, including detecting noise and reducing noise [8] [9] [10] [11] [2] [3]. There are two types of noise: attribute noise and class noise [9] [10]. Brodley [2] [3] detected and removed the class noise. Most of the work on noise detection is on supervised classification problems. However, Xiong et al. [8] studied the approach of remove noise for unsupervised learning. Kubica & Moore [11] 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 impacts of class noise on supervised learning algorithms and investigates the noise reduction 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 when the detecting and reducing noise approaches are applied to improve the data quality.

Improving the quality of the training data with mislabeled instances is important for supervised learning. Most of previous work focuses on removing the mislabeled instances on a specific learning algorithm – k-nearest neighbor (e.g., [3] [12] [13] and [14] created an instance selection mechanism for nearest neighbor classifiers). The algorithm proposed by [15] removes noise by retaining only those instances that have good impact on classifiers, which extends the nearest neighbor algorithm. Brodley and Friedl [3] identified and eliminated the mislabeled training instances by classifying each instance by an ensemble of classifiers. The ensemble of classifiers is built by three different learning algorithms (a 1-nearest neighbor, a linear machine, and a decision tree C4.5). An instance is removed if its original label is against the predictions of all three classifiers. Except the ensemble approach [3], the specific learning algorithm k-nearest neighbor is chosen for studying noise removing. To get under the noise impacts on different learning algorithms, we are going to investigate the performance of different supervised learning algorithms under the different noise levels, and their noise reduction sensitivity. It is expected to find which is better to be chosen among the popular supervised learning algorithms. We have done the investigation on four algorithms and will continue to investigate others.

Accuracy and AUC (area under the ROC curve) are two common measures for supervised learning algorithms. AUC is introduced into machine learning and data mining by [5] [6] from signal detection [4]. Bradley [16] indicates that AUC is more preferable to measure the performance of supervised learning algorithms. It discusses that AUC has several desirable properties compared to accuracy, such as increasing sensitivity in Analysis of Variance (ANOVA) tests, the independence of the decision threshold, and invariant to a priori class probability distributions. Ling [17] further compared the two measures with two formal criteria: (statistical) consistency, and (statistical) discriminancy. They formally prove that AUC is consistent with, and more discriminant (or finer) than accuracy, for the binary balanced datasets (which have the same number of positive and negative examples). In this paper, we compare AUC against accuracy under the noise situation. We investigate their noise reduction sensitivity and figure out which measure is preferable during improving data quality process.

3 Learning Algorithms

There exist more than 20 different learning algorithms, including the basic ones and improved variations. According to the categorization of WEKA [18], we choose the fundamental one from each category. That is, we are going to investigate the noise reduction 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) [19] 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 [20]). 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) [21] 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 for the real-world applications. Noise still impacts the performance of the model built on naïve bayes.

Support Vector Machine (SVM in short) [22] 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) [23], 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 Experiment Setup

To study the noise reduction 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 class noise). In order to have the datasets to study the noise reduction 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 it needs a wrong label, we 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 reduction sensitivity on the four learning algorithms step by step. We first focus on the impact of noise labels on binary classification. The features of the binary datasets from WEKA website are shown in Table 1.

TABLE 1
FEATURES OF THE 9 BINARY DATASETS USED IN THE EXPERIMENTS

	#Attributes	*#Examples*	*Class dist. (P/N)*
bmg	41	2417	547/1840
expedia	41	3125	417/2708
kr-vs-kp	37	3196	1669/1527
mushroom	22	8124	4208/3916
qvc	41	2152	386/1766
sick	30	3772	231/3541
spambase	58	4601	1813/2788
tic-tac-toe	10	958	332/626
travelocity	42	8598	1842/6756

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 features of the multiclass datasets from WEKA website are shown in Table 2.

TABLE 2
FEATURES OF THE 18 MULTICLASS CLASSIFICATION DATASETS USED IN THE EXPERIMENTS

	#Attributes	*#Examples*	*#Classes*
anneal	39	898	6
audiology	70	226	24
autos	26	205	7
balance	5	625	3
Glass	10	214	7
heart-c	14	303	5
heart-h	14	294	5
iris	5	105	3
letter	17	20000	26
lymph	19	148	4
primary-tumor	18	339	21
segment	20	2310	7
splice	62	3190	3
thyroid	30	3772	4
vehicle	19	846	4
vowel	14	990	11
waveform	41	5000	3
zoo	18	101	7

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 chosen 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. The performance of different learning algorithms is measured in accuracy and AUC. 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. The results are obtained from the implementation of WEKA for the four learning algorithms default parameter settings. That is, we use J48 for decision tree, NaiveBayes for naïve bayes, SMO for support vector machine, and Logistic for logistic regression [18].

5.1 Binary classification

First, we investigated the performance of the learning algorithms on binary classification datasets listed in Table 1. The accuracy and AUC of each algorithm on each dataset is shown in Table 3 and Table 5 respectively. We counted the number of champions (the highest result among the four algorithms, *in italic*) for each learning algorithm from Table 3 and Table 5 respectively and showed it in Table 4 and Table 6 (the number in the bracket means the tied champion). Table 4 shows that DT wins more than other algorithms, measured in accuracy. When the noise level is reduced, the number of wins for DT increases. Overall, DT performs the best, followed by Logistic, followed by SVM. NB performs the worst. Table 6 shows that Logistic wins more than other algorithms, measured in AUC. Logistic performs the best, followed by DT, followed by SVM. NB performs the worst. This shows that the two measures (accuracy and AUC) do not perform the same under the noise situation. Taking both measures into consideration, DT and Logistic are more preferable than SVM and NB.

Besides, in order to visualize the average performance of each learning algorithm under difference noise levels, we plotted the average accuracy and AUC in Figure 1 and Figure 2 respectively. Both figures show that 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. According to average accuracy, Figure 1 shows that NB performs the best when the noise level is the highest (50%), followed by DT and Logistic, followed by 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 reduction sensitive learning algorithms, DT performs the best, followed by Logistic, followed by SVM. In other words, if we have approaches which can improve the quality of labels, DT is preferable, followed by Logistic and SVM.

TABLE 3
THE ACCURACY OF THE FOUR LEARNING ALGORITHM ON BINARY CLASSIFICATION

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	*100.0*	*100.0*
	NB	*96.14*	96.71	97.32	97.56	97.21	95.87
	SVM	74.09	*97.89*	*99.86*	*99.98*	99.98	*100.0*
	Logistic	74.63	96.07	99.74	99.93	99.98	*100.0*
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 4
NUMBER OF CHAMPIONS OF THE FOUR LEARNING ALGORITHM ON THE NINE BINARY CLASSIFICATION DATASETS (ACCURACY)

	%Noise	50%	40%	30%	20%	10%	0%
#Wins	DT	4	5	5	6	7	6+(1)
	NB	3	1	0	0	0	0
	SVM	0	2	2	2	1	1+(1)
	Logistic	2	1	2	1	1	1+(1)

5.2 Multiclass classification

We further study the noise reduction sensitivity of the four learning algorithms in multiclass classifications. Note that the implementation (SMO and Logistic) for support vector machine and logistic regression respectively can be directly applied to multiclass classifications. We have done the experiments on 18 datasets listed in Table 2. Because of the space limitation, we plotted the figures, instead of showing the results in Tables which occupy more space. Figure 3 shows the accuracy of each learning algorithm on each dataset, and

Figure 4 shows the AUC. We summarize the performance of the four learning algorithms by counting the champions and show the number of champions in Table 7 (in accuracy) and Table 8 (in AUC). Besides, we show the average performance of the four learning algorithms in Figure 5 (average accuracy) and Figure 6 (average AUC).

TABLE 5

THE AUC OF THE FOUR LEARNING ALGORITHM ON BINARY CLASSIFICATION

Dataset	%Noise	50%	40%	30%	20%	10%	0%
bmg	DT	0.705	*0.758*	*0.783*	*0.816*	*0.835*	*0.851*
	NB	0.723	0.723	0.726	0.729	0.732	0.731
	SVM	0.500	0.500	0.500	0.500	0.500	0.503
	Logistic	*0.767*	0.757	0.769	0.771	0.774	0.776
expedia	DT	0.521	0.544	0.782	0.826	0.865	0.877
	NB	0.805	0.814	0.821	0.822	0.821	0.822
	SVM	0.515	0.527	0.566	0.655	0.687	0.726
	Logistic	*0.904*	*0.904*	*0.909*	*0.911*	*0.912*	*0.914*
kr-vs-kp	DT	0.943	0.955	*0.982*	*0.991*	*0.996*	*0.997*
	NB	0.928	0.928	0.930	0.934	0.938	0.948
	SVM	0.592	0.839	0.908	0.909	0.941	0.954
	Logistic	*0.970*	*0.976*	0.980	0.984	0.990	0.996
mushroom	DT	0.970	0.987	0.997	0.999	1.000	1.000
	NB	0.997	0.997	0.997	0.997	0.998	0.998
	SVM	0.750	0.980	0.999	1.000	1.000	1.000
	Logistic	*1.000*	*1.000*	*1.000*	*1.000*	*1.000*	*1.000*
qvc	DT	0.687	0.768	0.794	0.822	*0.844*	*0.867*
	NB	0.733	0.745	0.745	0.746	0.744	0.746
	SVM	0.500	0.500	0.505	0.525	0.542	0.547
	Logistic	*0.812*	*0.818*	*0.823*	*0.827*	0.828	0.831
sick	DT	0.848	0.897	0.906	0.891	0.926	0.930
	NB	*0.884*	0.902	0.905	0.915	0.915	0.923
	SVM	0.575	0.694	0.504	0.500	0.500	0.500
	Logistic	0.880	*0.910*	*0.917*	*0.932*	*0.937*	*0.933*
spambase	DT	0.893	0.891	0.898	0.908	0.919	0.934
	NB	0.922	0.927	0.929	0.931	0.936	0.937
	SVM	0.524	0.67	0.883	0.897	0.895	0.893
	Logistic	*0.950*	*0.953*	*0.956*	*0.959*	*0.964*	*0.972*
tic-tac-toe	DT	0.500	0.641	0.745	0.794	0.840	0.895
	NB	0.739	0.758	0.756	0.76	0.762	0.764
	SVM	0.504	0.938	0.948	0.966	0.974	0.974
	Logistic	*0.983*	*0.981*	*0.980*	*0.984*	*0.992*	*0.997*
travelocity	DT	0.927	0.959	0.978	*0.991*	*0.995*	*0.995*
	NB	0.934	0.932	0.925	0.911	0.920	0.918
	SVM	0.544	0.579	0.558	0.556	0.557	0.563
	Logistic	*0.972*	*0.980*	*0.984*	0.979	0.986	0.972
average	**DT**	**0.777**	**0.822**	**0.874**	**0.893**	**0.913**	**0.927**
	NB	**0.852**	**0.858**	**0.859**	**0.861**	**0.863**	**0.865**
	SVM	**0.566**	**0.692**	**0.708**	**0.723**	**0.733**	**0.740**
	Logistic	***0.915***	***0.920***	***0.924***	***0.927***	***0.931***	***0.932***

TABLE 6

NUMBER OF CHAMPIONS OF THE FOUR LEARNING ALGORITHM ON THE NINE BINARY CLASSIFICATION DATASETS (AUC)

	%Noise	50%	40%	30%	20%	10%	0%
#Wins	DT	0	1	2	3	3+(1)	3+(1)
	NB	1	0	0	0	0	0
	SVM	0	0	0	(1)	(1)	(1)
	Logistic	8	8	7	6+(1)	6+(1)	6+(1)

Figure 5 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 high (50%). Although it performs better when the noise level reduces, the performance of the other three learning algorithms improves more significantly. When the noise level is higher than 20%, SVM performs the best in average, followed by DT, followed by Logistic. However, when the noise level is further reduced, DT performs the best, followed by SVM, followed by Logistic. Table 7 also shows that DT wins most champions, especially after the noise level is reduced to less than 30%. Surprisingly, NB wins more champions than SVM and Logistic, especially when the noise level is not lower than 40%.

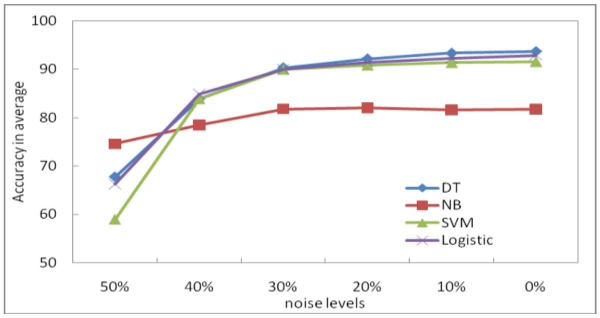

Figure 1: The average accuracy over the nine binary datasets for the four learning algorithms.

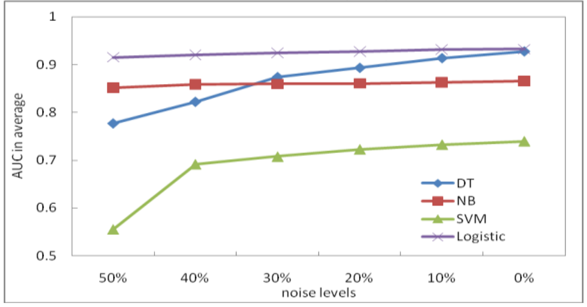

Figure 2: The average AUC over the nine binary datasets for the four learning algorithms.

TABLE 7

NUMBER OF CHAMPIONS OF THE FOUR LEARNING ALGORITHM ON THE 18 MULTICLASS CLASSIFICATION DATASETS (ACCURACY)

	%Noise	50%	40%	30%	20%	10%	0%
#Wins	DT	4	4	5	8	10	6
	NB	6	6	4	5	5	2
	SVM	4	3	6	3	2	3
	Logistic	4	5	3	2	1	7

The experimental results in AUC are shown in Figure 4, with the average AUC shown in Figure 6. The summarization of the number of champions of each learning algorithm in AUC is displayed in Table 8.

TABLE 8

NUMBER OF CHAMPIONS OF THE FOUR LEARNING ALGORITHM ON THE 18 MULTICLASS CLASSIFICATION DATASETS (AUC)

	%Noise	50%	40%	30%	20%	10%	0%
#Wins	DT	0	0	1	2	2	2
	NB	10	9	9	8	6	6
	SVM	1	2	2	1	1	1
	Logistic	7	7	6	7	9	9

Surprisingly, the experimental results in AUC show us different observations. NB wins most champions showing in Table 8. When the noise level is lower than 20%, Logistic wins most champions. DT and SVM only win a few. Figure 6 also shows that NB dominates other in AUC, followed by Logistic, SVM, and DT. Again, the different observations from accuracy and AUC further confirm that the two measures perform differently in noise situations. The surprising observation based on AUC motivates us to further study NB and Logistic under noise in the future. It can be expected that we can improve the accuracy of NB and Logistic by adjusting the classification threshold, instead of using the classical 0.5. We also observe that SVM performs well in accuracy on binary and multiclass classification, but its AUC is very low. What is the reason? We are going to investigate this and to further improve its performance in AUC.

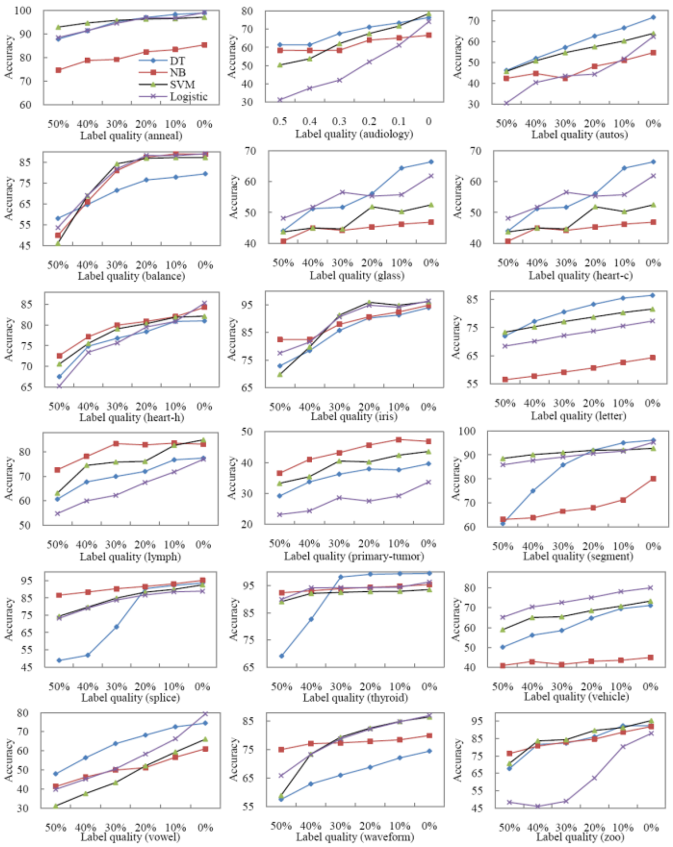

FIGURE 3. THE ACCURACY OVER THE 18 MULTICLASS DATASETS FOR THE FOUR LEARNING ALGORITHMS.

6 Conclusions and Future Work

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 of each algorithm depends on the percentage of noise injected and the characteristics of different datasets.

We also investigated the noise reduction sensitivity of the two measures (accuracy and AUC). It is observed that AUC is more noise tolerant. The improvement of AUC is slower with the noise reduction. When we study the noise reduction (quality improvement), accuracy is the preferable measure.

This study is very useful in situations of real-world data processing that may contain implicit and explicit errors. The results in accuracy show that naïve bayes is the most noise tolerant algorithm. However, decision tree performs the best overall under different noise level for most datasets (binary and multiclass), followed by logistic regression and support vector machine. However, logistic regression performs the best in AUC. When we develop approaches to improve the data quality (reduce the noise level), they are more preferred than naïve bayes. Besides, we are also interested in studying how to improve the accuracy performance of naïve bayes, as it performs well in AUC; and how to improve the AUC performance of support vector machine, as it performs well in accuracy.

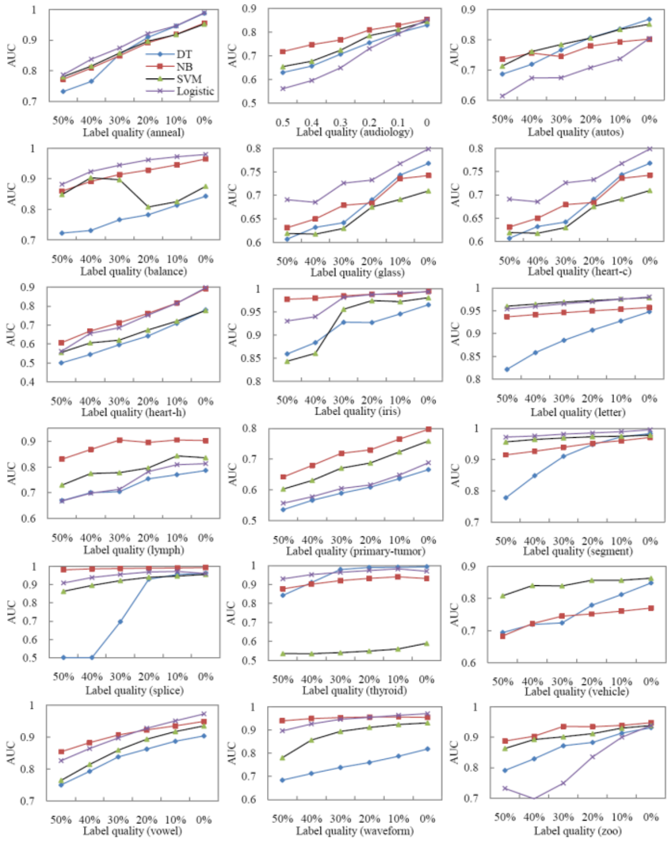

FIGURE 4. THE AUC OVER THE 18 MULTICLASS DATASETS FOR THE FOUR LEARNING ALGORITHMS.

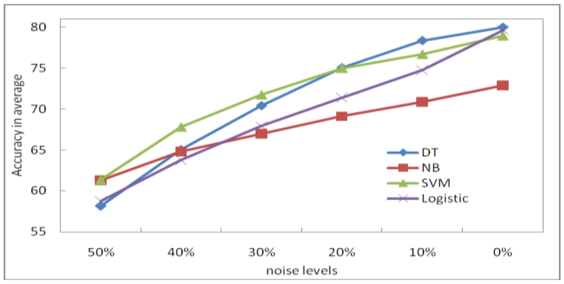

Figure 5. The average accuracy over the 18 multiclass datasets for the four learning algorithms.

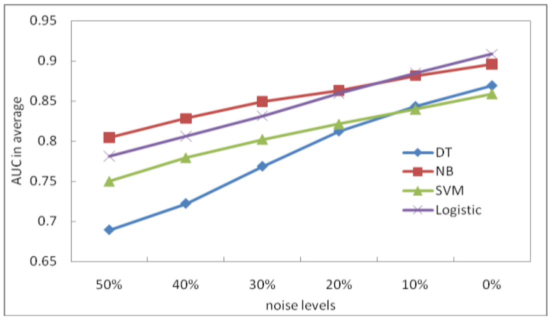

Figure 6. The average AUC over the 18 multiclass datasets for the four learning algorithms.

7 Acknowledgment

We thank the anonymous reviewers for the valuable comments. The work was supported by the National Science Foundation (IIS-1115417).

8 References

[1] Sheng, V.S., Provost, F. and Ipeirotis, P. Get another label? Improving data quality and data mining using multiple, noisy labelers. Proceedings of the Fourteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2008), 2008, 614-622.

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

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

[4] D.M. Green and J.A. Swets. Signal Detection Theory and Psychophysics. Wiley, New York, 1966.

[5] F. Provost and T. Fawcett. Analysis and visualization of classifier performance: comparison under imprecise class and cost distribution. Proceedings of the Third International Conference on Knowledge Discovery and Data Mining, AAAI Press, 1997, 43-48.

[6] F. Provost, T. Fawcett, and R. Kohavi. The case against accuracy estimation for comparing induction algorithms. In Proceedings of the Fifteenth International Conference on Machine Learning, Morgan Kaufmann, 1998, 445-453.

[7] J. Huang and C.X. Ling. Constructing New and Better Evaluation Measures for Machine Learning. The Twentieth International Joint Conference on Artificial Intelligence (IJCAI 2007), 859-864.

[8] 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.

[9] 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, 920-927.

[10] 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.

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

[12] Gamberger, D.; Lavrac, N.; and Dzeroski, S. 1996. Noise elimination in inductive concept learning: A case study in medical diagnosis. In In Proc. of the 7th International Workshop on Algorithmic Learning Theory, 199-212. Springer.

[13] Wilson, D. R., and Martinez, T. R. 2000. Reduction techniques for instance-basedlearning algorithms. Journal of Machine Learning 38:257-286.

[14] Skalak, D. B. 1994. Prototype and feature selection by sampling and random mutation hill climbing algorithms. In Proceedings of the Eleventh International Conference on Machine Learning, 293-301.

[15] Aha, D. W.; Kibler, D.; and Albert, M. K. 1991. Instance based learning algorithms. Journal of Machine Learning 6:37-66.

[16] A. P. Bradley. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognition, 30:1145-1159, 1997.

[17] C. X. Ling, J. Huang, and H. Zhang. AUC: a statistically consistent and more discriminating measure than accuracy. In Proceedings of 18th International Conference on Artificial Intelligence (IJCAI-2003), 2003, 519-526.

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

[19] J.R. Quinlan. C4.5: Programs for Machine Learning. Morgan Kaufmann: San Mateo, CA, 1993.

[20] Tom Mitchell, Machine Learning, McGraw Hill, 1997

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

[22] Corinna Cortes and V. Vapnik, Support-Vector Networks, Machine Learning, 20, 1995.

[23] Hilbe, Joseph M. Logistic Regression Models. Chapman & Hall/CRC Press, 2009.