Evaluation of Class Noise Impact on Performance of Machine Learning Algorithms

Zahra Nazari†, Masooma Nazari†, Mir Sayed Shah Danish††, and Dongshik Kang†††

†Graduate School of Engineering & Science, University of the Ryukyus, Okinawa 903-0213, Japan
††Department of Electronics & Electrical Engineering, University of the Ryukyus, Okinawa 903-0213, Japan
†††Department of Information Engineering, University of the Ryukyus, Okinawa 903-0213, Japan

Summary
Real-world datasets are not perfect and always suffer from noise that may affect classifiers built under the effect of such type of disturbance. Different types of noise are existing in almost any real-world problem, but not always known. Existence of noise decreases the accuracy of a classifier and increases its training time and complexity of the induced model. Most of existing machine learning algorithms have integrated different approaches to enhance their learning abilities in presence of noise, but noise still can make negative impacts. Therefore noise robustness of a classifier is an important issue in noisy environments and should be studied. This paper evaluates the robustness of different machine learning algorithms against class noise. The Equalized Loss of Accuracy (ELA) is the robustness metric which is used in this study. Ten benchmark datasets with 0-20% of noise level are used in experiments and finally ELA results of algorithms are compared.

Key words:
Noise impact, Classification, Robustness metric, Class noise

1. Introduction

Most of real-world datasets suffer from noise or corruptions due to malfunctions, unfortunate calibrations of measurement equipment or network problems during the transport of sensor information to a central measurement collection unit, etc. This issue adversely effects on performance of classifiers since quality of training data has direct influence on classifier accuracy. The problem of learning in noisy environments has been the focus of many researches in machine learning and data mining and most of learning algorithms have integrated various approaches to enhance their learning abilities from noisy environment. But presence of class noise still can have negative impact on performance of learning algorithms. However the overall performance of an algorithm depends not only on the data quality, but also on the proportionality of the classification algorithm which is used for data [1, 2]. Therefore it's very important to know what kinds of classification algorithms are more appropriate for working with noisy data.

One of the common ways to know which algorithm is more suitable to deal with noisy data is to check the accuracy of classifiers over a fixed collection of datasets regardless of the noise level present in the data. However this procedure is not enough, since the study of performance alone cannot provide enough information about classifier behavior against noise. Accordingly, a study with a controlled noise level for each dataset is required to achieve a meaningful conclusion while evaluating classifier behavior in noisy environment [1, 2, 4]. In addition to focus on classic performance assessment, robustness of the method in noisy environment also should be studied. Performance is defined as the accuracy of a classifier to predict the label of a new instance and noise. Robustness is of an algorithm is defined as the capability to build models that are not sensitive to data corruptions and suffer less from the impact of noise. A robust classification algorithm can build models from noisy data which are more similar to models built from clean data. Finally combination of the robustness and performance concepts can make a unified conclusion on the expected behavior of the method in the noisy environment [1].

There are some robustness measures which are introduced during the past years. The robustness measure introduced by Kharin et al. (1994) considers the performance of Bayesian Decision rule as a reference, which is considered as the classifier providing the minimal risk when the training data are not corrupted [5]. Relative Loss of Accuracy (RLA) introduced by Sáez et al. (2011) evaluates the robustness as the loss of accuracy with respect to the case without noise [6]. The Equalized Loss of Accuracy (ELA) is the correction of RLA measure which is also introduced by Sáez et al. (2016). ELA combines the robustness and a factor depending on the initial accuracy. This measure tries to minimize the problems of considering performance and robustness measure individually and can be used to easily compare different classifiers [1]. In this paper we use ELA measure to evaluate the three widely used machine learning algorithms namely Decision Learning Tree (DLT)

Manuscript received August 5, 2018.
Manuscript revised August 20, 2018.

[7], Support Vector Machines (SVM) [8], and K Nearest Neighbors (KNN) [9]. Ten benchmark datasets from repository of University of the California, Irvine (UCI) are used in our experiments [10].

The rest of this paper is organized as follows. In section 2 we will present an introduction about data quality and noise and importance of data quality in learning. Section 3 describes what algorithm robustness measurement is. The ELA measure will also be explained in detail in this section. In section 4 we present experimental results and finally we conclude this paper in section 5.

2. Data Quality and Noise

There are many components that determine the quality of any dataset, but commonly the quality of a dataset can be characterized by attributes and class labels. The quality of attribute represents how well attributes describe an instance for the learning purpose and quality of class label indicates whether the label of each instance is correctly assigned or not [2]. To train a classifier, usually a set of attributes will be selected to characterize the class label with two assumptions: 1) there is a correlation between attributes and class; it is clear that correlations of some attributes with the class are stronger than others, thus they play more important roles in classification. 2) there is weak interaction between attributes; therefore the learning algorithm consider each attributes independently to induce the classifier. Though, real-world data rarely comply with the above assumptions and usually they contain some attributes with very little correlations with the class or the interactions among attributes are very strong [2, 3].

Accordingly, the quality of a dataset is determined by internal and external factors. Internal factor indicates whether attributes and class are well selected or not; external factors indicate errors into attributes and class labels. Both internal and external factors are used to define noisy instances, where noise is anything that destructs the relationship between attributes and class. Noise negatively affects the system performance in terms of classification accuracy, size, time in building and interpretability of the model obtained. There are three types of major physical sources of noise: 1) inadequate description for attributes or/and class; 2) corrupted attribute values in the training examples; and 3) incorrect classification of training examples. Since it is difficult to characterize the adequacy of the description for attribute and class in real-world data, only last two sources are considered. Therefore the physical sources of noise in machine learning can be categorized into attribute noise and class noise [2, 11].

2.1. Class Noise

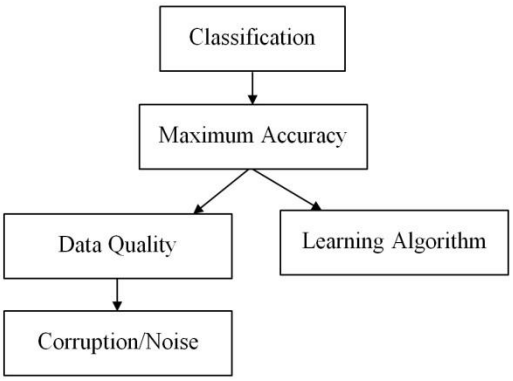

Fig 1. Data quality in classification problem [12].

Class noise is also known as labelling errors occurs when an incorrect label assigns to an instance. There are many studies that have been done to deal with class noise and most of them have suggested that in many situations, elimination of instances with class noise will increase the classification accuracy [1, 2, 13]. There are two possible sources for class noise:

1) Contradictory instances: some instances appear more than once in the dataset but with different labels.

2) Misclassification: some instances are labeled incorrectly; this problem occurs when different classes have similar symptoms.

2.2. Attribute Noise

Corruption/error represented in the attribute values of instances in a dataset is called attribute noise. Erroneous attribute values, missing or don't know attribute values, and incomplete attributes or don't care values are some causes for attribute noise. There are some research efforts have been done to deal with attribute noise and elimination is one of the common ways which is used in these studies. However their results show eliminating instances containing attribute noise is not a good idea, because many other attributes of an instance may contain valuable information [2, 13, 15]. Therefore, handling attribute noise is more difficult and research in this area has not made much progress, except some efforts on handling missing attribute values popularized by Cohen and Cohen [16].

In this paper, the most common and most disruptive type of class noise known as *misclassification* is considered. Misclassification refers to those instances which are labeled incorrectly.

3. Algorithm Robustness Measurement

Noise are common in real-world datasets and prevent knowledge extraction from the data and models which are obtained using these noisy data are spoiled comparing to the models learned from the clean data. Therefore, learning algorithms with high robustness against noise are desired. Robustness of an algorithm is defined as the capability to build models that are not sensitive to data corruptions and suffer less from the impact of noise. A robust classification algorithm can build models from noisy data which are more similar to models built from clean data. However high robustness of a classifier is not enough to judge about its behavior with noisy data, because a good behavior implies a high robustness but also a high performance of classifier without noise [1, 4]. There are some robustness measures which are used to analyze the degree of robustness of the classifiers. One of the measures introduced by Kharin el al. considers the performance of Bayesian Decision rule as a reference, which is considered as the classifier providing the minimal risk when the training data are not corrupted [5].

$$BRM_{x\%} = \frac{E_{x\%} - E}{E} \quad (1)$$

where $E_{x\%}$ is the risk of the classifier with the x% noise level, and E is the risk of the Bayesian Decision rule without noise. RLA is another measure which evaluates the robustness as the loss of accuracy with respect to the case without noise [6]. This method evaluates the robustness as the loss of accuracy with respect to the case without noise. The ELA is the correction of RLA measure which is used in this paper and explained in details in section 3.1.

$$RLA_{x\%} = \frac{A_{0\%} - A_{x\%}}{A_{0\%}} \quad (2)$$

where $A_{0\%}$ is the accuracy of the classifier with 0% noise level and $A_{x\%}$ is the accuracy of classifier with x% noise level.

3.1. The ELA Measure

This measure is proposed by Sáez et al. and combines the robustness and a factor depending on the initial accuracy. This measure tries to minimize the problems of considering performance and robustness measure individually and can be used to easily compare different classifiers [1]. ELA takes into account the noiseless performance when considering which classifier is more appropriate with noisy data and this makes ELA more suitable to compare the behavior or different classifiers against noise. The ELA measure is:

$$ELA_{x\%} = \frac{100 - A_{x\%}}{A_{0\%}} \quad (3)$$

where $A_{0\%}$ is the performance without noise and $A_{x\%}$ is the performance at a noise level x%.

$ELA_{x\%}$is the measure of behavior with noise at a given noise level x% which is based on: 1) robustness of the method, which is the loss of performance at a controlled noise level x%; and 2) the behavior with noise for the clean data, that is without controlled noise ($ELA_{0\%}$).

4. Experiments and Results

In this section we present the way the experiments were conducted. DLT, SVM with Gaussian kernel and KNN are three machine learning algorithms which are used for classification in this study. Ten benchmark datasets from UCI repository are used for the experiments. In the following Table 1 summarizes the evaluated classification algorithms, Table 2 presents characteristics of all datasets and Fig 2 shows how to introduce noise into the dataset.

Table 1. A summary of evaluated machine learning algorithms in the noisy environment.

Algorithm	Type	Description
DLT	Classification	Uses a decision tree as a predictive model.
KNN	Regression/ Classification	Classifies an instance by majority votes of its neighbors.
SVM	Classification	Builds a model that predicts whether a new example falls into one category or the other.

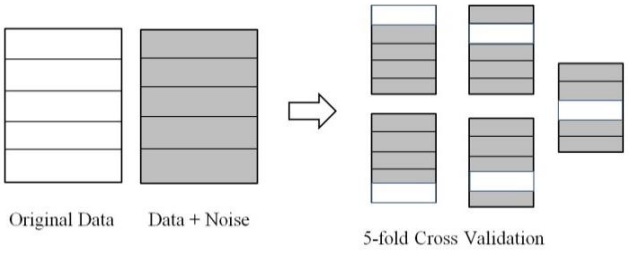

Fig 2. Noise introduction into the dataset [12].

Table 2. Characteristics of datasets used in experiment.

Dataset	Number of Attributes	Number of Classes	Number of Instances
Appendicitis	7	2	106
Banknote authentication	5	2	1372
Breast cancer	9	2	286
Glass	10	6	214
Heart disease	14	5	303
Iris flowers	4	3	150
Ionosphere	34	2	351
Liver disorder	7	2	345
Pima	9	3	768
Thyroid gland	6	3	215

In order to be able to control the noise level in a dataset, we manually added noise into each training dataset. According to *random class noise scheme* [2], we add *x*% noise into a dataset by randomly changing the class labels of exactly *x*% of the instances by other one out of the other classes. As mentioned before, we will evaluate machine learning algorithms with benchmark datasets with *x*=0%, *x*=10% and *x*=20% noise added to each dataset. The classification accuracy of three algorithms will be computed on the 30 datasets with and without noise along with their corresponding ELA result for each noise level. All parameters of learning algorithms are selected by heuristic search and classifiers are trained in 5-fold cross validation form. The number of neighbor in KNN is set to 10 and maximum number of splits in DLT is set to 20. The classification accuracy of mentioned algorithms with 0% noise, 10% and 20% are presented in Tables 3 to 5 respectively.

Table 3. Accuracy of classifiers with noise level of 0%.

Dataset	Classification Accuracy		
	DLT	SVM	KNN
Appendicitis	84%	**87%**	**87%**
Banknote authentication	98%	**100%**	**100%**
Breast cancer	68%	**74%**	73%
Glass	**68%**	**68%**	66%
Heart disease	51%	60%	**61%**
Iris flowers	95%	**96%**	95%
Ionosphere	87%	**94%**	84%
Liver disorder	64%	**69%**	64%
Pima	74%	**75%**	73%
Thyroid gland	90%	**96%**	91%

Table 4. Accuracy of classifiers with noise level of 10%.

Dataset	Classification Accuracy		
	DLT	SVM	KNN
Appendicitis	79%	83%	**85%**
Banknote authentication	96%	**99%**	**99%**
Breast cancer	71%	**74%**	70%
Glass	60%	**65%**	62%
Heart disease	48%	56%	**57%**
Iris flowers	93%	**95%**	94%
Ionosphere	83%	**89%**	81%
Liver disorder	60%	**67%**	62%
Pima	69%	**73%**	70%
Thyroid gland	86%	**93%**	89%

Table 5. Accuracy of classifiers with noise level of 20%.

Dataset	Classification Accuracy		
	DLT	SVM	KNN
Appendicitis	77%	81%	**82%**
Banknote authentication	92%	**99%**	97%
Breast cancer	68%	**73%**	70%
Glass	60%	**65%**	59%
Heart disease	47%	**55%**	53%
Iris flowers	91%	**93%**	92%
Ionosphere	82%	**89%**	80%
Liver disorder	58%	**66%**	62%
Pima	64%	**73%**	69%
Thyroid gland	84%	**92%**	87%

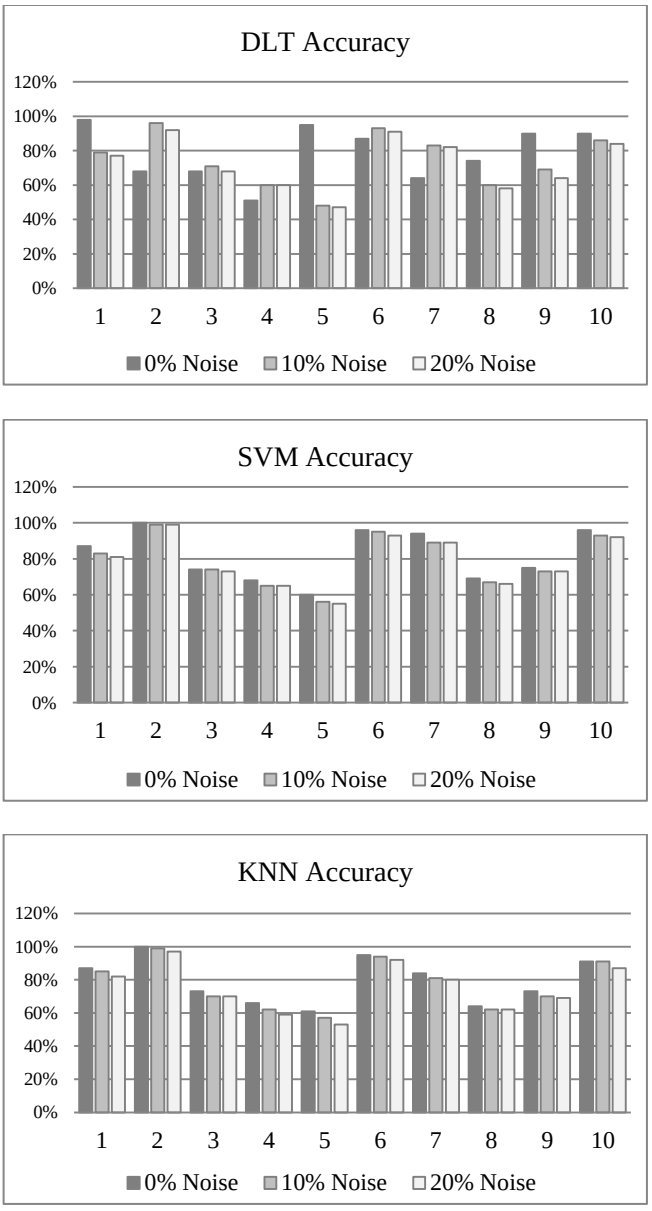

Fig 3. The above graphs present the performance of LDT, SVM and KNN classifiers for 10 datasets. (1. Appendicitis, 2. Banknote, 3. Breast cancer, 4. Glass, 5. Heart disease, 6. Iris, 7. Ionosphere, 8. Liver disorder, 9. Pima, 10. Thyroid gland).

To evaluate the impact of class noise on machine learning algorithm performance, we have executed our experiments on the 10 datasets (described in Table 2) with various levels of added class noise. DLT, SVM and KNN are trained by these noisy datasets and impacts of class noise on their accuracy are presented. Considering the above tables and fig 3, we found that existence of class noise decreases the classification accuracy and without these noisy instances classification accuracy is better. The ELA results at 10% and 20% noise for all datasets are presented in below tables (Table 6 and 7).

Table 6. The ELA result at **10%** noise level. Best results are remarked in bold.

Dataset	ELA Result at 10% Noise		
	DLT	SVM	KNN
Appendicitis	0.2469	0.1945	**0.1622**
Banknote authentication	0.0424	**0.0037**	0.0059
Breast cancer	0.4305	**0.3516**	0.4053
Glass	0.5919	**0.5026**	0.5813
Heart disease	1.0105	0.7298	**0.7073**
Iris flowers	0.0736	**0.0520**	0.0631
Ionosphere	0.1954	**0.1170**	0.2261
Liver disorder	0.625	**0.4782**	0.5937
Pima	0.4189	**0.36**	0.4109
Thyroid gland	0.1555	0.0729	**0.0689**

Table 7. The ELA result at **20%** noise level. Best results are remarked in bold.

Dataset	ELA Result at 20% Noise		
	DLT	SVM	KNN
Appendicitis	0.2778	0.2137	**0.2065**
Banknote authentication	0.0825	**0.0066**	0.0234
Breast cancer	0.4773	**0.3654**	0.3854
Glass	0.5783	**0.5091**	0.6160
Heart disease	1.0366	**0.7525**	0.7681
Iris flowers	0.0947	**0.0729**	0.0842
Ionosphere	0.2068	**0.1170**	0.2380
Liver disorder	0.6562	**0.4927**	0.5937
Pima	0.4864	0.36	**0.2246**
Thyroid gland	0.1777	**0.0833**	0.1428

Regarding the performance results presented in Table 3 to 5 and Fig 3, SVM with Gaussian kernel has a better performance with and without noise than DLT and KNN classifiers. Considering ELA results presented in Table 6

and 7, the robustness of SVM against noise is better than two other algorithms. In case of 10% noise level, the ELA result of SVM is better for 7/10 datasets and in case of 20% noise level is better for 8/10 datasets.

5. Conclusion

In this study we evaluated the class noise impact on the performance of three widely used machine learning algorithms namely DLT, SVM and KNN. Different levels of class noise 10% and 20% have been added into the original dataset and they are classified in 5-fold cross validation form. In addition to classification accuracy, the ELA metric is also used to measure the robustness of each algorithm against noise. Ten datasets from UCI repository with different number of attributes and instances are used in our experiments and SVM classifier presented better performance and robustness against noise than other classifiers for most of these datasets with and without added noise.

Our future work is to increase the number of machine learning algorithms to evaluate as well as number of datasets. Impact of attribute noise is another important topic that should be evaluated and compared with impact of class noise.

Acknowledgment

The authors would like to express their cordial thanks to University of the Ryukyus.

References

[1] J. A. Sáez, J. Luengo, and F. Herrera, "Evaluating the classifier behavior with noisy data considering performance and robustness: The Equalized Loss of Accuracy measure" ELSEVIER, Neurocomputing 176 (2016) 26-35.

[2] X. Zhu and X Wu, "Class Noise vs. Attribute Noise: A Quantitative Study of Their Impacts" Kulwer Academic Publishers, Artificial Intelligence Review 22:177-210, 2004.

[3] S. Kim, H. Zhang, R. Wu and L. Gong, "Dealing with Noise in Defect Prediction" 33rd International conference on software engineering, May 21-28, 2011, Waikiki, Honolulu, HI, USA, ACM 978-1-4503-0445-0/11/05.

[4] E. Kalapanidas, N. Avouris, Marian Craciun and D. Neagu, "Machine Learning Algorithms: a Study on Noise Sensitivity" Balcan conference in informatics, 2003.

[5] Y. Kharin, E. Zhuk, "Robustness in statistical pattern recognition under contaminations of training samples" In proceeding of the 12th IAPR International Conference on Pattern Recognition, Conference B: Computer Vision and Image Processing, Vol. 2, 1194, pp. 504-506.

[6] J. A. Sáez, J. Luengo, and F. Herrera, "Fuzzy rule based classification systems versus crisp robust learners trained in presence of class noises effects: a case of study" In 11th International Conference on Intelligent Systems Design and Applications (ISDA 2011), 2011, pp.1229-1234.

[7] J. R. Quinlan, "Introduction of Decision Trees" Kluwer Academic Publisher, Machine Learning 1:81-106, 1986.

[8] Z. Nazari and D. Kang, "Density Based Support Vector Machines for Classification", International Journal of Advanced Research in Artificial Intelligence, vol.4, No. 4 2015.

[9] Z. Zhang, Introduction to machine learning: k-nearest neighbors, Ann. Transl.Med. 4 (11) (2016) 1–7.

[10] Dua, D. and Karra Taniskidou, E. (2017). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[11] L. P. F. Garcia, A. C. P. F. Carvalho and A. C. Lorena "Effect of label noise in the complexity of the classification problems" Neurocomputing 160 (2015) 108-119, ELSEVIER

[12] http://sci2s.ugr.es/noisydata

[13] L. P. F. Garcia, A. C. Lorena and A. C. P. F. Carvalho, " A Study on Class Noise Detection and Elimination" Brazilian Symposium on Neural Networks (SBRN), IEEE, 2012, pp. 13–18.

[14] X. Zhu, X. Wu and Q. Chen, "Eliminating Class Noise in Large Datasets" In proceeding of 20th International Conference on Machine Learning (ICML 2003), Washington DC.

[15] C. M. Teng, "Correcting Noisy Data", In proceeding of 16th international conference on Machine Learning (ICML), pp. 239-248.

[16] J. Cohen and P. Cohen, "Applied multiple regression correlation analysis for the behavioral science (2nd ed.), Hillsdale, NJ:Erlbaum