Study and Analysis of Noise Effect on Big Data Analytics

Sushma Rani N
Asst Professor, MVGRCE (A)
Research Scholar, CUTM
sushmaranin@mvgrce.edu.in

Dr P.Srinivasa Rao
Associate Professor, MVGRCE (A)
psr.sri@gmail.com

Prof. Anurag
Pro Vice Chancellor
CUTM
anurag@cutm.ac.in

Abstract- Noise in training set can lead to poor accuracy of the classifier. Due to this the results may be misleading in assigning the class label. Due to this the number of training samples increase which leads to complexity in analysis of the data. A great about of research is done on the study of noise in data set and various methodologies were proposed on how to handle noise. This paper gives a survey on the various techniques to handle noise data and experiment results showing how noise effects on various evaluation metric is given.

Keywords: Class noise, Attribute noise, Uniform Class noise, ensemble classifiers

1. Introduction:

With advancements in technologies and emerging trends the raw data that is available is abundant. With the evolving big data techniques the quantity as well as the variety of data has increased [2]. As data is gathered from various sources and received in variety of formats the data has to be pre-processed to aid as an input for machine learning algorithms. The raw data has many anomalies and unwanted information which can be known as noise. Noise data can degrade the accuracy of the machine learning algorithm. Noise cannot be avoided and induces errors; this affects the efficiency of the algorithm. In real time applications like in the field of medicine, satellite communication, stock market for example the accuracy of the algorithm plays a very vital role. With noise data present in the data set the misclassification error rate increases which may lead to critical issues. For estimation and detection of noise the traditional machine learning techniques does not give accurate results for massive datasets, so many techniques were proposed to detect [1], estimate noise [6] and eliminate noise so that the efficiency of the algorithm increases.

Noise has two main sources [8] in which implicit errors occur as they are introduced due to measurement tools like sensors; random errors are due to batch processes systems and when

the data is gathered by experts. Due to these reasons the quality of data is estimated by 2 factors, external factor and internal factors: the internal factor gives details whether the selection and definition of the class and the attributes are done to characterize the underlying theory, and the external factor represent the errors introduced into the class and the attribute.

The main effects of noise data is that it reduces the classification accuracy, increases the classification model building time, increases the size of the classifier and interpretability of the classifier.

Considering the above scenario there are 3 types of major physical sources of noise: 1.Insufficiency of the description for attributes or the class (or both); 2.corruption of attribute values in the training examples; and 3.erroneous classification of training examples [8].

Noise data can be defined as corrupted data and will occur due to 2 types of errors: implicit errors and random errors. With many observations done on the data the sources of noise in can be distinguished into 2 types 1. Attribute noise; and 2.Class noise.

The two types of noise considered are class and attribute noise, have been modeled using four different noise schemes which are categorized into two as:

Class Noise: Uniform class noise and Pairwise class noise.

Attribute Noise: Uniform attributes noise and Gaussian attributes noise.

Hence the data collected from real-world problems tend to be imperfect and often suffer from data that is corrupted which may affect and hinder the model performance in terms of [9] the accuracy, model building time, size of the classifier and interpretability of the classifier. If the dataset used to train the model is affected and corrupted due to noise, both the learning phase and the model obtained will be negatively affected [14].

As per various experiment results attribute noise tend to be more harmful and also has been stated [7] that attribute noise is harmful in the attributes that are highly correlated with the class label.

In machine learning algorithms identifying class noise is crucial [6]. Class noise can exist for different reasons and in various types of applications. The taxonomy of class noise was generated based on the work of [13, 17, 18, 19, and 20]. The various sources of labeled noise i.e. class noise is mentioned and its taxonomy is given in [16].

This paper focuses on a survey of the type of noises that occur in real world dataset and the techniques to detect class noise estimate and eliminate class noise. It also deals with experimental results on how noise effects the various evaluations metric on 3 different machine learning classification algorithms. Two datasets –Diabetic and Titanic datasets are used for experimentation.

The paper is further organized as section 2 deals with the related work, section 3 gives the methodologies used in dealing with noise, section 4 deals with the hardware and software requirements, section 5 deals with the experimental results and comparisons of various machine learning algorithms used for classification and section 5 gives the conclusion and future work.

2. Related Work

Class noise can be due different reasons for example in disease prediction as data comes from experimental values, food labeling, natural language processing etc. [6] suggested the basic 2 strategies to handle class noise: 1.How to do learning with class noise, 2. How to eliminate Class noise.

Many strategies were proposed to eliminate noise. Methods were also proposed to detect and estimate noise. Elimination techniques attempts to eliminate the samples with high noise probabilities and eliminate from the training set.

Many strategies used ensemble classifier for handling noise. Ferhat Ozgur Catak [2] explained about practical multiple ensemble classifier training models to classify large- scale datasets. In big data traditional classification algorithms cannot scale up to the size of the data. So a data partitioning strategy is adopted for training high dimensional data. In this the noise filtering approach the one-class SVM algorithm is applied to handle and remove noise instances. The future enhancements that can be made as suggested in the work as to study different noise removing methods to clean sub data set and adaptive noise removing ratio to make the method as autonomous.

José A. Sáez et al [7] aims to develop a good analysis of the behavior of Multiple Classifier Systems (MCSs) with noisy data with respect to their individual components. The hypothesis about the behavior of MCSs with noisy data will be checked in detail and the conditions under which the MCSs studied work well with noisy data will be analyzed. The experiments were conducted on large collection of real-world datasets and different types of noise and various noise levels were introduced to draw meaningful conclusions. In this work the performance and robustness of the classifiers were compared.

Methods were proposed to detect and eliminate class noise as [1] presented a global architecture for Class noise detection and elimination in large datasets. The architecture initially partitions the data into subsets. It then extracts association rules from each set later applying classifiers. Finally the all the results are combined to obtain the final decision. An ensemble of classifier were used in [3] which is a simplistic noise handling strategies for classification datasets that use ensembles of classifiers for noise identification is proposed. The classifiers recommended and used are SVM, k-nn, CART, C4.5, Random Forest, Naïve Bayes, Multi-layer Perceptron. Experiments were performed using various combinations of classifiers and suggested that the model can be extended for the investigation strategies for multi class datasets. Class noise detection techniques can be categorized into graph based model or classification based model [6].

To reduce the effect of noise, filtering techniques can be used. Jose A. Saez [14] mentioned that to reduce the effect of noise on the classifier two approaches were followed in the work. 1. Algorithm level approach and 2. Data level approach. The paper also suggested that data level approaches are more popular and are independent of the classifier. A new noise filtering technique was proposed “Iterative Noise Filter based on the Fusion of Classifiers (INFFC)” based on various noises elimination filtering techniques.

Noise can also be estimated. Lin Gui et al [6] proposed a novel method for class noise estimation and learning with noise strategies. The method improves the learning performance for both online and offline learning algorithms. The authors further suggested it can be extended to semi-supervised learning algorithms. Rakshita Pandya, Kshitij Pathak [11] proposed noise estimation method and noise removal using Support Vector Machine (SVM), which used Non-parametric noise estimator and Practical selection of meta-parameters for SVM regression. Lei Han et al [12] proposed a method to efficiently deal with the large class problem by paying attention to a small subset of candidate classes instead of the entire class space. Given a data point x (without observing y), we select a small number of competitive candidates as the estimation is referred to as Candidates vs. Noises Estimation (CANE). We show that CANE is consistent and its computation using stochastic gradient method is independent of the class size K.

Class labeling may always not be binary so with non-binary classification [4] proposed a method which uses adaptation or development of new techniques for handling class noise within non-binary classification paradigm which can be extended to hybrid methods to reduce incorrect filtering.

Techniques were also proposed in imbalanced classification by re-sampling methods with filtering. José A. Sáez et al [5] focuses on minority class which is most interesting from the application point of view, but tends to be misclassified by standard classifiers. This work focuses on studying the influence of noise and borderline examples in generalization of The Synthetic Minority Over-sampling Technique (SMOTE).

Bootkrajang, J., Kabán [10] stated the problem of multi-class classification in the presence of labeling errors was studied. The authors proposed a generative multi-class classifier to learn with labeling errors, which extends the multi-class quadratic normal discriminant analysis by a model of the mislabeling process. They demonstrated the benefits of this approach in terms of parameter recovery as well as improved classification performance.

3. Methodology

Simulating the noise of real-world datasets

The initial amount of noise and the type of noise present in the data set are not known; hence an assumption cannot be made on the type of the noise and level of noise. Hence an assumption is made that the datasets used are free from noise. In order to predict the effect of noise on the classification algorithm and to predict the efficiency of the algorithm noise can be introduced into each dataset.

The 3 main characteristics for characterizing Noise generation are: [7]

1. Where the noise is introduced.
2. The distribution of noise.
3. The magnitude of generated noise values.

The two types of noise the class noise and attribute noise can be modeled using 4 different schemes.

1. Class noise : It can be modeled using
   A. Uniform class noise: Here x% of the original data is corrupted.
   B. Pairwise class noise. If considered X is the majority class and Y the second majority class, an example with the label X has a probability of X/100 of being incorrectly labeled as Y.
2. Attribute noise
   A. Uniform attribute noise: x% of the values of each attribute in the dataset are corrupted.
   B. Gaussians attribute noise: Attribute values are corrupted, adding a random value to the attributed.

The Taxonomy of Class Noise Model

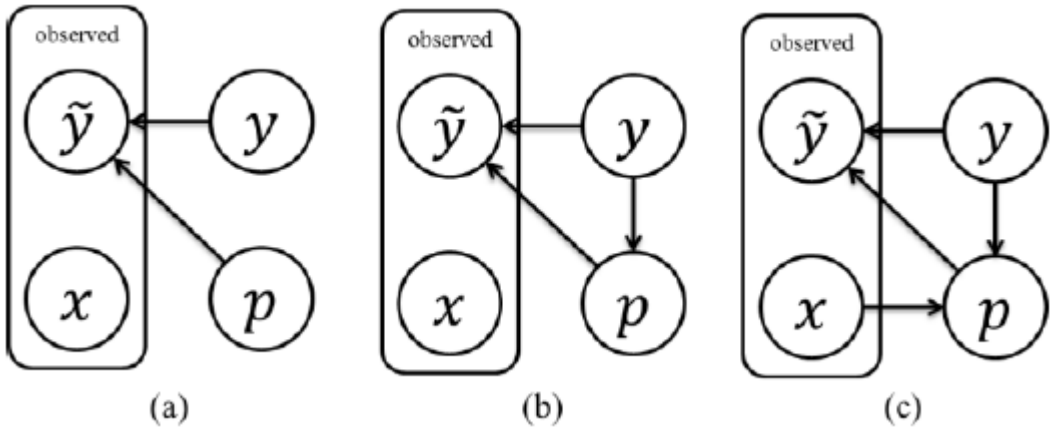


Figure 1: The taxonomy of class noise

Class noises can be categorized into three different models based the dependence of noise to y, where y is the true label of the sample of x and p where p is the class noise rate [15]. The first model showed in Figure.1 (a) Noise Completely at Random Model [13]. In the model the class noise rate is completely random and independent of the labels and the feature set. The second model shown in Figure.1(b) Noise at Random Model [17, 18]. In the model, the class noise rate is dependent on the true label of a sample and independent of the feature set of a sample. The third model shown in Figure.1 (c) Noise not at Random Model [19, 20]. This model assumes the class noise rate should be affected by both the label and the feature set of the sample.

The procedure is divided into three steps. [3] The first steps involves that several classifiers are induced for the complete training data. This step is known as ensemble of classifiers. The ensemble generally comprises of smallest odd number of the classifier set.

The classifiers used in this work are K-nearest neighbor (K-nn), Support Vector Machine (SVM), Decision Tree classification algorithms.

The fundamental definitions of noise rate, minimization of loss function and risk are stated in [6].

According to [6] it is assumed that let the distribution of clean data be D, and (x1,y1), (x2,y2),…, (xn,yn) denote n training samples from D with true binary label yi (yi=±1,i=1,2,…,n). Let the distribution with class noise be $\tilde{D}$, where the label $\tilde{y}i$ may be different from the true label yi.

Definition.1:

In a given distribution the definition of class noise rate is given as the probability of the observed label different from the true label of xi, denoted by $P(\tilde{y} = y|x)$

As given in the definition the rate of class noise estimation has to be done for each sample and prior knowledge is not available. Hence an estimation method is required to estimate the class noise rate. A good class noise estimation method can be adopted as Support Vector Machine as proposed in [11].

Definition.2:

In this the real-value (R) decision function is define as $\boldsymbol{(x) = P\,(y=1|x) - 1/2}$.

The risk function is given by $R_D\ (f) = E_{(x,y)\sim D}\ (1_{sign(f(x)\neq y)})$.

The loss function is $l(f(x), y)$ with a real-value prediction, for the clean distribution where y=±1 is the true label of x and the loss function on the noisy distribution with an observed label, denoted as $\tilde{l}(f(x),\tilde{y})$.

There are three different types of risks – empirical, expected risk with the noisy distribution, expected risk under clean distribution.

According to Definition 1 we need to estimate the class noise rate but this cannot be done as the data is noisy. The data can be modeled by using K-nearest neighbor. Hence the noise data can be replaced with nearest neighbor value.

The machine learning classification algorithms used for the experimentation are: k nearest neighbor (K-nn), Support Vector Machine (SVM) and Decision Tree. The various evaluating metric used to compare the classifiers are Accuracy, Precision, Recall, F-score, Sensitivity and Specificity.

Experiment was conducted by evaluating the classifiers with the real world data taken from two different datasets: Diabetics and Titanic with no noise induced. Later Uniform class noise is introduced into the dataset with different percentage i.e. 5% and 20% and the classifiers are evaluated. K-nn impute is used to fill in missing values which further enhances the accuracy of the classifier.

4. Environmental Setup

The experiments were conducted on a system configured with 2.5GHz Processor, 8 GB RAM and 1 TB storage space installed with Anaconda studio, Python. The evaluation of model is carried out by using various quality metrics such as Precision, Recall, F-Score and Accuracy.

5. Results and Discussion

The results are evaluated in three scenarios

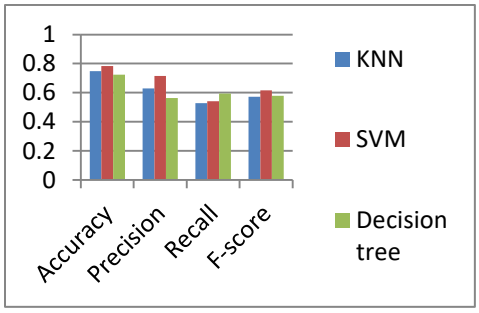
- Without Noise induction
- Noise induced – 5% and 20%
- Noise induced – 5% and 20% and performing k-nn impute

The algorithms are compared in all the above scenarios using evaluation metric – Accuracy, Precision, Recall, and F-score.

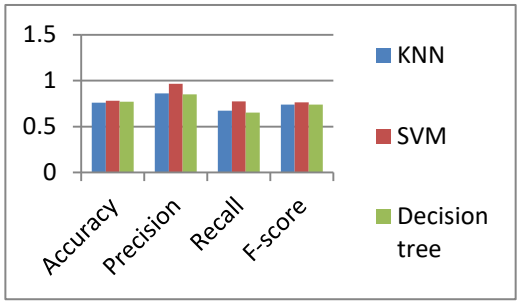
The various classifiers K-nearest neighbor (K-nn), Support Vector Machine (SVM), Decision Tree on which experiment was conducted are compared in all the three scenarios.

The results show that the classifiers outperform when noise was induced and imputed using K-nn impute. The accuracy also varied with the percentage of noise induced.

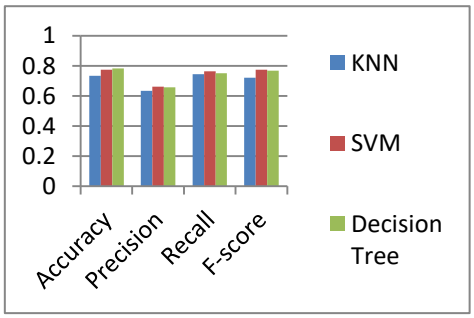
Comparative results using various evaluating metric



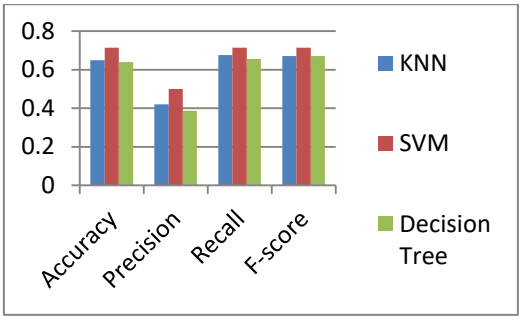
Graph 4.1 Diabetes Dataset with no induced noise



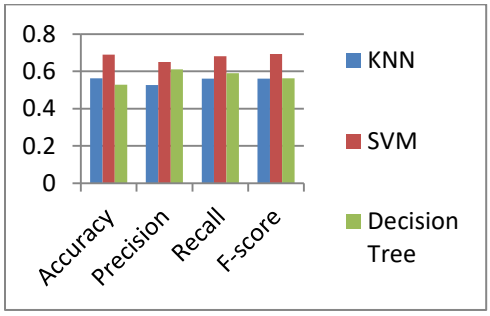
Graph 4.2 Titanic Dataset with no induced noise



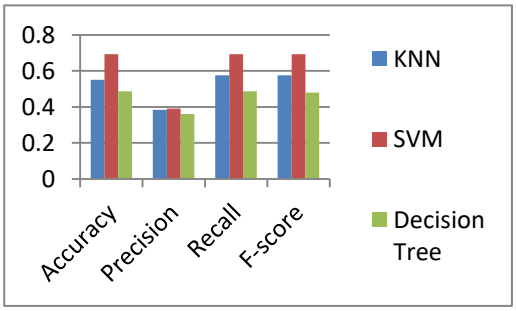
Graph 4.3 Diabetes Dataset with 5 % induced noise



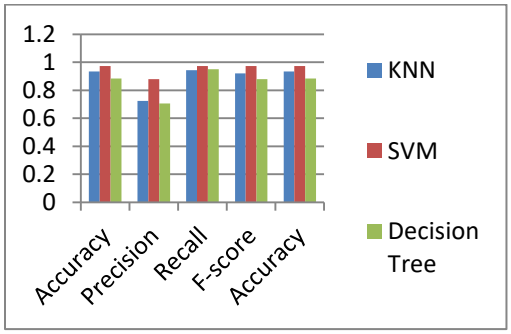
Graph 4.4 Titanic Dataset with 5% induced noise

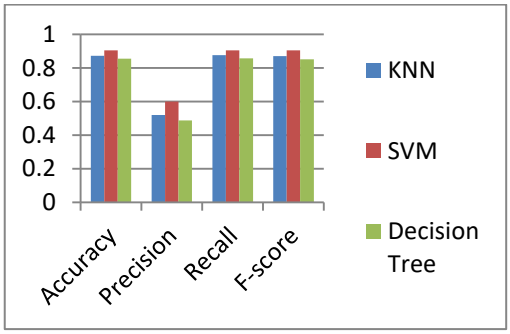


Graph 4.5 Diabetes Data Set with 20 % induced noise

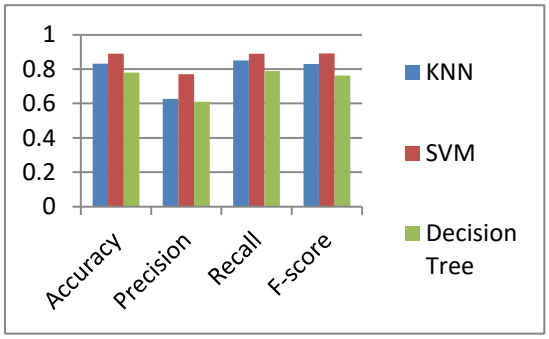


Graph 4 .6 Titanic Dataset with 20% induced noise

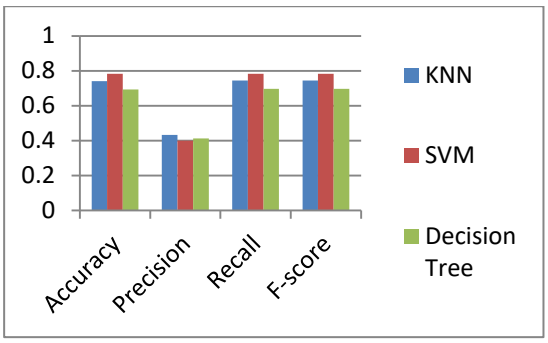




Graph 4.7 Diabetes Dataset with 5 % induced noise and k-nn impute



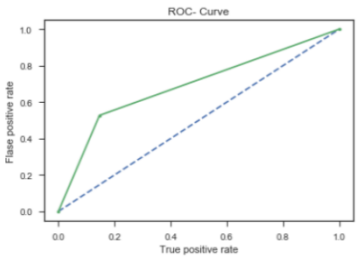
Graph 4.8 Titanic Data Set with 5% induced noise and k-nn impute



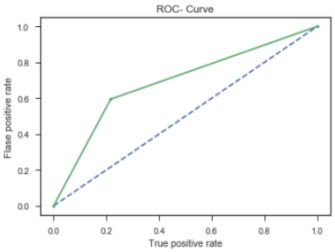
Graph 4.9 Diabetes Dataset with 20 % induced noise and k-nn impute

Graph 4.10Titanic Data Set with 20% induced noise and k-nn impute

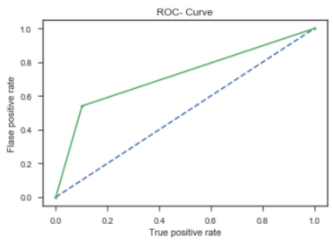
ROC Curves: DIABETES WITH NO INDUCED NOISE



Graph 4.11KNN

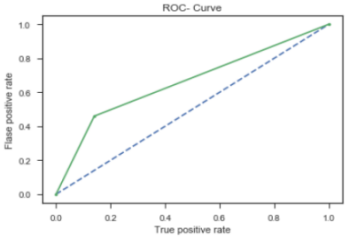


Graph 4.12 SVM

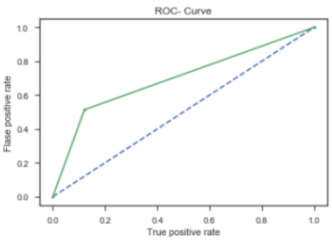


Graph 4.13 DECISION TREE

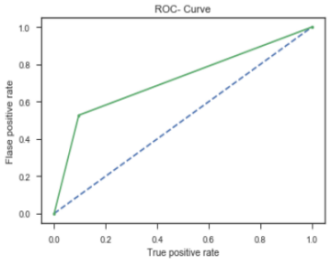
MISSING VALUE DATASETS: (5% MISSING) - WITHOUT IMPUTATION



Graph 4.14KNN

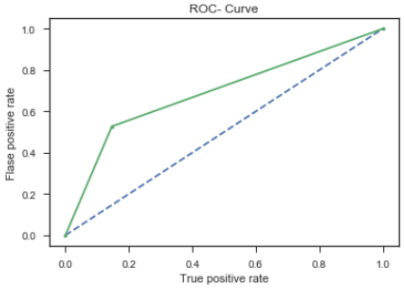


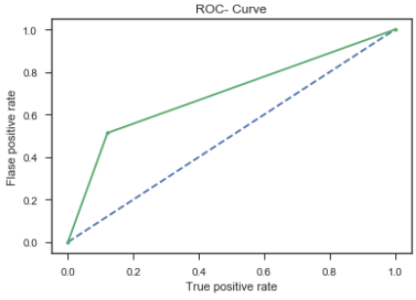
Graph 4.15 SVM

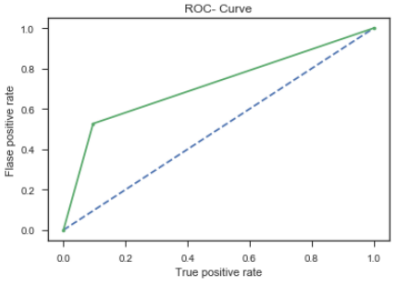


Graph 4.16 DECISION TREE

MISSING VALUE DATASETS: (5% MISSING) - WITH IMPUTATION METHOD KNN-IMPUTE







Graph 4.17KNN Graph 4.18 SVM Graph 4.19 DECISION TREE

5. Conclusion and Future Work

The paper focused on various noises that hinder the performance of machine learning classification algorithm. It focused on various types of noises, the sources of noise and their effect on the classifier. The paper showed the how the various evaluation metric vary with noise-no noise induced, noised induced with different percentages, noise induced and imputed. The work can be extended by using different machine learning algorithms like semi –supervised and deep learning techniques to improve the accuracy of the dataset with noise of different types, induced at various levels.

References:

[1] Btissam Zerhari1 et al, "Detection and Elimination of Class Noise In Large Datasets Using Partitioning Filter Technique", 4th IEEE International Colloquium On Information Science and Technology (Cist) IEEE, 2016

[2] Ferhat Ozgur Catak, "Robust Ensemble Classifier Combination Based On Noise Removal with One-Class SVM" , ICONIP 2015 Springer International Publishing Switzerland 2015, Part II, LNCS 9490, Pp. 10–17, 2015.

[3] Lu´Is Paulo F. Garcia et al," A Study on Class Noise Detection and Elimination", 2012 Brazilian Symposium on Neural Networks, 20-25 Oct. 2012.

[4] Ronaldo C. Prati et al, "Emerging Topics and Challenges of Learning From Noisy Data In Nonstandard Classification: A Survey Beyond Binary Class Noise", Knowledge And Information Systems, Springer-Verlag London Ltd., Part Of Springer Nature 2018.

[5] José A. Sáez et al, "SMOTE–IPF: Addressing the Noisy and Borderline Examples Problem In Imbalanced Classification By A Re-Sampling Method With Filtering", Information Sciences 291, 184-203,2015.

[6] Lin Gui et al, "A Novel Class Noise Estimation Method And Application In Classification" CIKM '15 Proceedings Of The 24th ACM International On Conference On Information And Knowledge Management, 1081-1090, Melbourne, Australia, October 18 - 23, 2015.

[7] José A. Sáez et al, "Tackling The Problem Of Classification With Noisy Data Using Multiple Classifier Systems: Analysis of the Performance and Robustness", Information Sciences 247, 1-20, 2013.

[8] X. Zhu, X. Wu, "Class Noise vs. Attribute Noise: A Quantitative Study", Artificial Intelligence Review 22, 177-210, 2004.

[9] X. Wu, X. Zhu, "Mining With Noise Knowledge: Error-Aware Data Mining", IEEE Transactions on Systems, Man, And Cybernetics 38, 917-932, 2008.

[10] Bootkrajang, J., Kabán,' A.: Multi-Class Classification In The Presence Of Labeling Errors. In: European Symposium On Artificial Neural Networks 2011 (ESANN 2011), 345- 350, 2011

[11] Rakshita Pandya, Kshitij Pathak, "Survey On Noise Estimation And Removal Methods Through SVM", International Journal Of Computer Applications (0975 – 8887) Volume 86 No 9, January 2014.

[12] Lei Han et al, "Candidates vs. Noises Estimation for Large Multi-Class Classification Problem", Cornel University Library, Statistics, Machine Learning, ICML, 2018.

[13] Angluin, D and D.Laird, P. "Learning From Noisy Examples." In Machine Learning 2(4): 343-370, 1988.

[14] Jose A. Saez, "INFFC: An Iterative Class Noise Filter Based On the Fusion Of Classifiers with Noise Sensitivity Control", Information Fusion 27, 19-32, 2016.

[15] Frénay, B., and Verleysen, M... "Classification In The Presence Of Label Noise: A Survey". In IEEE Transactions On Neural Networks And Learning Systems, Vol. 25, 5, 2014.

[16] Benoît Frénay And Michel Verleysen, "Classification In The Presence Of Label Noise: A Survey", IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, VOL. 25, NO. 5, MAY 2014.

[17] Lawrence, N. D., And Schölkopf, B... "Estimating A Kernel Fisher Discriminant In The Presence Of Label Noise," In Proceeding Of International Conference On Machine Learning 06–313, 2001.

[18] Perez, C. J., Giron, F. J., Martin, J., Ruiz, M., and Rojano, C... "Misclassified Multinomial Data: A Bayesian Approach," Revista De La Real Academia De Ciencias Exactas Físicas Y Naturale Serie A Matemáticas, Vol. 101, No. 1, 71-80, 2007.

[19] Kolcz, A., and Cormack, G. V... "Genre-Based Decomposition of Email Class Noise," In Proceeding of 15th ACM SIGKDD Conference On Knowledge Discovery And Data Mining, 427–436, 2009.

[20] Zhu, X., Wu, X., And Chen, Q. J... "Eliminating Class Noise In Large Datasets." In Proceeding of International Conference on Machine Learning, Vol. 3, 920-927. 2003.