Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 161 (2019) 466–474

Procedia Computer Science

www.elsevier.com/locate/procedia

The Fifth Information Systems International Conference 2019

Dealing with Noise Problem in Machine Learning Data-sets: A Systematic Review

Shivani Gupta[a,*], Atul Gupta[b]

[a]Manipal University Jaipur, India

[b]Indian Institute of Information Technology, Design and Manufacturing Jabalpur, India

Abstract

The occurrences of noisy data in data set can significantly impact prediction of any meaningful information. Many empirical studies have shown that noise in data set dramatically led to decreased classification accuracy and poor prediction results. Therefore, the problem of identifying and handling noise in prediction application has drawn considerable attention over past many years. In our study, we performed a systematic literature review of noise identification and handling studies published in various conferences and journals between January 1993 to July 2018. We have identified 79 primary studies are of noise identification and noise handling techniques. After investigating these studies, we found that among the noise identification schemes, the accuracy of identification of noisy instances by using ensemble-based techniques are better than other techniques. But regarding efficiency, usually single based techniques method is better; it is more suitable for noisy data sets. Among noise handling techniques, polishing techniques generally improve classification accuracy than filtering and robust techniques, but it introduced some errors in the data sets.

Keywords: Noise; Class noise; Attribute noise; Types of noise; Noise identification techniques; Noise handling techniques; Classification

1. Introduction

Machine learning (ML) and data mining (DM) is the process of finding useful results from real world data sets. The real world data contains irrelevant or meaningless data termed as noise which can significantly affect various data analysis tasks of machine learning are classification, clustering and association analysis. The need to address this noise is clear as it is detrimental to almost any kind of data analysis. We may have two types of noise in machine learning dataset: in the predictive attributes(attribute noise) and the target attribute (class noise). The presence of noise in a data set can increase the model complexity and time of learning which degrades the performance of learning algorithms. Therefore, there is a need to identify and handle these noise in data sets.

* Corresponding author. Tel.: +91-940-658-0983
E-mail address: shivani.gupta@jaipur.manipal.edu

10.1016/j.procs.2019.11.146

Table 1. Research questions.

RQ#	Research questions
RQ1	How we classify the noise in data sets?
RQ2	Which ML techniques has been used for identification of class noise and attribute noise?
RQ3	Which ML techniques has been used for handling of class noise and attribute noise?

This SLR aims to identify and analyze the techniques used to handle noise in data sets in studies published over 17 years (between January 1993 and July 2018). In this paper, we aim to review a number of approaches to identify and handle noise in data sets by following Kitchenhams [1]. To perform a systematic review, we divide the paper into two parts according to the noise present in the data sets: class noise and attribute noise. In practice, existing empirical studies use different techniques to identify and handle the class noise and attribute noise in data sets.

This paper is organized as follows. In the next section, we provide our systematic literature review methodology. Section 3 presents the different types of noise present in classification data sets. Section 4 contains the different techniques used for identification of noise in data sets. In section 5 we present the different techniques used to handle noise in data sets. Our results of systematic review is presented in Section 6. We concludes the study and future directions in Section 7.

2. Methodology

A systematic approach to applied to reviewing the literature on the identification and handling class noise in data sets. We follow the systematic review approach suggested by Kitchenhams [1].

Our systematic review followed the steps outline below, some of which with iteration:

- Identification of the need of the review.
- Formulation of the research questions.
- Identification of relevant literature by conducting a comprehensive and exhaustive search.
- Selection of primary studies based on inclusive/exclusive criteria.
- Data extraction together with the quality assessment.
- Interpretation of results.

2.1. Research questions

This systematic mapping review (SLR) aims to analyze the techniques used to identify and handling noise in classification data sets. Towards this aim, six research questions (RQs) were raised as follows given below in Table 1.

2.2. Search strategy and study selection

We formed the refined search terms by combining alternative terms and synonyms using Boolean expression OR and AND. We considered journal papers, conference proceedings, workshops, symposiums, and ACM/IEEE bulletins to conduct the searches. Keyword searching was performed on following electronic databases as we considered these to be the most relevant ones: ACM Digital Library, IEEE Xplore, Science Direct, SpringerLink, Google Scholar. After performing an initial search, the relevant studies were determined by obtaining the full-text papers following the inclusion and exclusion criteria described in the next section. We included the empirical studies using the different ML techniques for identifying and handling noise in classification data sets.

2.2.1. Inclusion criteria

We defined the following inclusion criteria, a study must be reported in a complete paper published in English. We have included empirical studies which based on noisy instances, mislabelled instances, class noise, label noise, attribute noise, techniques used for identification of noise, and techniques used for handling of noise.

2.2.2. Exclusion criteria

The paper must not be on Non-empirical studies, Studies included imbalanced data sets, Studies by same author in conference as well extended version in journal, i.e, similar, and Review studies.

2.2.3. Identification and classification of papers

Included studies were published between January 1999 and July 2018. The main three key elements to our searches: manual reading of paper titles, keyword searching using search engines and manually scanned the references list of each relevant study. The goal of classification is divided into two-fold: first to review the recent studies on identification of noise and second to review the recent studies on handling of noise.

We plotted to further present the distribution of research attention in each publication year in Fig. 1. There are some studies which are included in both noise identification and handling.

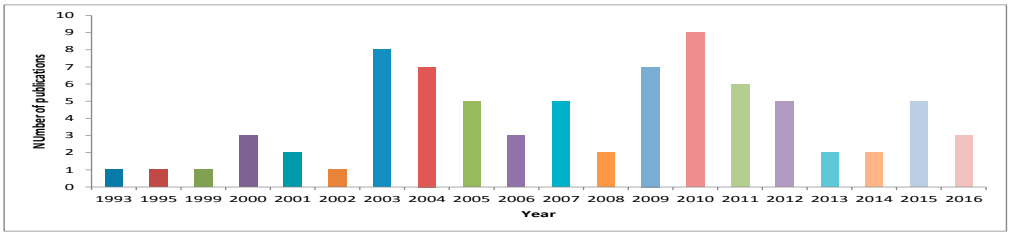

Fig. 1. No. of paper included per year.

2.3. Quality assessment

The quality of a study may constitute a criterion form inclusion, exclusion or supporting data analysis and synthesis. The criteria are based on four quality assessment (QA) questions which addressing to RQ4: The details of the studies

Table 2. Quality assessment questions.

Q#	Quality questions
Yes	Partly
No	
QA1	Is the paper based on research or is it merely a lessons learned report based on expert opinion?
QA2	Was the data collected in a way that addressed the research issue?
QA3	Is the study of value for research or practice?
QA4	Is there a clear statement of findings?
QA5	How estimate the impact of that study?
QA6	Does the study have adequate number of the average citation count per year?
QA7	Is there any comparative analysis conducted ?
QA8	Is the data set size appropriate?
QA9	Are the performance measures used to assess the classifier models clearly defined?

addressing each specific research questions are provided in Table 2. The scoring procedure was Y = ✓, N = ✗. Finally, after thorough reviews and discussions a final decision was made to the inclusion/exclusion for each study.

3. Results and Discussion

This section presents and discusses the findings of this review in response to our classification scheme of noise identification and handling techniques.

3.1. Description of primary studies

A summary of the selected studies quality is presented in Table 3 with author name, year and reference.

Table 3. Selected primary studies.

ID	Author	Year	Ref
N1	D. Gamberger and S. Dzeroski	2000	[3]
N2	Rebbapragada, Umaa, and Carla E. Brodley	2007	[4]
N3	C. Q. Zhu X, Wu X	2003	[5]
N4	X. Zeng and T. Martinez	2003	[6]
N5	D. G. Borut Sluban and N. Lavra	2010	[8]
N6	Q. C. Xingquan Zhu, Xindong Wu	2010	[9]
N7	X. Zhu and X. Wu	2006	[10]
N8	L. Libralon and A. C. Lorena	2009	[11]
N9	S. D. N. Segata, E. Blanzieri	2009	[12]
N10	C. Libralon, Giampaolo L. and A. C. Lorena	2009	[13]
N11	S. Kim, H. Zhang and L. Gong	2011	[14]
N12	R. Moosavi, M. Fazaeli Javan	2010	[15]
N13	S. Z. Huawen Liua	2012	[16]
N14	L.K. Soh and J. Bernadt	2003	[17]
N15	A. C. A. Miranda, L. Garcia and A. Lorena	2009	[18]
N16	M. Prem Melville, Nishit Shah	2004	[19]
N17	N. L. Borut Sluban, Dragan Gamberger	2012	[20]
N18	A. C. L. Garcia, Luis Paulo F. and A. C. Carvalho	2012	[21]
N19	M. R. Smith and T. Martinez	2011	[22]
N20	W. T. Zhong, Shi and T. M. Khoshgoftaar	2005	[23]
N21	J. V. H. Khoshgoftaar, Taghi M. and A. Napolitano	2011	[24]
N22	W. Jeatrakul, Piyasak and C. C. Fung	2010	[25]
N23	L. Guan, Donghai and S. Lee	2011	[26]
N24	Zhang, Peng, Xingquan Zhu, Yong Shi, Li Guo, and X. Wu	2011	[2]
N25	C.-M. Teng	2001	[27]
N26	T. M. K.-J. V. H. Folleco, Andres and L. Bullard	2008	[28]
N27	C. M. Teng	2004	[29]
N28	W. Sun, Jiang-wen and S. fu Chen	2007	[30]
N29	V. Khoshgoftaar, T	2005	[31]
N30	J. Kubica and A. Moore	2003	[32]
N31	C. M. Teng	2003	[33]
N32	B. Yan Zhang, Xingquan Zhu	2005	[34]
N33	S. Gernot Armin Liebchen, Bhekisipho Twala	2007	[35]
N34	S. Shah and A. Kusiak	2010	[36]
N35	C. M. Teng	2003	[37]
N36	X. Zeng and T. R	2001	[38]
N37	S. L. Muhlenbach, Fabrice and D. A. Zighed	2004	[39]
N38	D. G. Sluban, Borut and N. Lavra	2010	[40]
N39	Cao, Jingjing, Sam Kwong, and Ran Wang	2012	[41]
N40	Lorena, Ana C., and A. C. P. L. F. Carvalho	2004	[72]
N41	S. Ipeirotis, Panagiotis G. and J. Wang	2013	[42]
N42	C. M. Teng	2000	[43]
N43	F. M. Lallich, Stphane and D. A. Zighed	2002	[44]
N44	Smith, Michael R., and Tony Martinez	2014	[45]
N45	Dong Wang, Xiaoyang Tan	2014	[46]
N46	Fefilatyev, D., Kasturi, R. and Bunke, H.	2012	[47]
N47	Jose A. Saez, Mike Galar, Julian Luengo, Francisco Herrera	2016	[48]
N48	Jose A.Saez, Julian Luengo , Francisco Herrera	2016	[49]
N49	Luis P.F.Garcia, Andre C.P.L.F.deCarvalho , AnaC.Lorena	2016	[50]
N50	Garcia, LPF, de Carvalho ACPLF, Lorena AC	2015	[51]
N51	Cesa-Bianchi and Shamir, Ohad	2010	[52]
N52	Borut Sluban, Nada Lavrac	2015	[53]
N53	Maryam Sabzevari, Gonzalo Martinez-Munoz, Alberto Suarez	2015	[54]
N54	Pelletier, Charlotte et al	2017	[73]
N55	Yuan, Weiwei, Donghai Guan, Tinghuai Ma et al	2018	[74]
AN1	Taghi M. Khoshgoftaar	2005	[55]
AN2	Ling Sun, Jia-Yu Chi, Zhong-Fei LI1	2006	[56]
AN3	Michael Mannino, Yanjuan Yang , Young Ryu	2009	[57]
AN4	Jiye Li and Nick Cercone	2000	[58]
AN5	Khoshgoftaar, Taghi M and Van Hulse, Jason	2005	[59]
AN6	Kamal M. Ali Michael J. Pazzani	1993	[60]
AN7	Jason D. Van Hulse Taghi M Khoshgoftaar and Haiying Huang	2007	[61]
AN8	Wah, Catherine and Belongie, Serge	2010	[62]
AN9	Taghi M. Khoshgoftaar	2004	[63]
AN10	Omid Naghash Almas and Wan Mei Tang	2016	[64]
AN11	Michael Mannino, Yanjuan Yang and Young Ryu	2009	[65]
AN12	Goldman Sally A. and Robert H. Sloan	1995	[66]
AN13	Khoshgoftaar, Taghi M., and Jason Van Hulse	2009	[67]
AN14	Ying Yang, Xindong Wu, and Xingquan Zhu	2004	[68]
AN15	Lukaszewski, Tomasz, et al.	2011	[69]
AN16	Hua Yin, Hongbin Dong, Yuxuan Li	2009	[70]

3.2. Publication source

The major publications were in IEEE Transactions on Machine learning, Journal of Pattern recognition, IEEE Transactions on Software Engineering, Machine learning Research, Journal of Pattern recognition, Pattern recognition letters, Data and Knowledge Engineering, Information Sciences and so on. The studies included in our investigation 59% were present in journals and 41% were present in conferences.

3.2.1. Quality assessment

A summary of the questions used to assess the quality of these studies is presented in Section II. The score for each study is shown in Table 4.

3.2.2. Impact of included studies

We can estimate the impact of included studies, by counting the number of times a study has been cited. Figure 2 shows citation counts of included studies.

As we can see, the lowest citation count and the highest citation counts are 0 and 487, respectively. Forty-two papers (around 81%) have a citation count in the range of 0-50, only ten papers (19 %) have high citation counts in the range of 50-147.

3.3. RQ1: How we classify the noise in classification problems?

The real world data set can usually be characterized by two information sources: (1) attributes, and (2) class labels. Depending on these sources there are two types of noise present on the classification problems: class noise and attribute noise.

3.3.1. Class noise

The class labels noise represents whether the class of each instances is correctly assigned or not. There are two possible sources for class noise:

1. Contradictory instances: The same instances appear more than once in the data set and are labeled with different class labels. For example the same instances with different class labels.

Table 4. Quality Assessment.

Paper ID	QA1	QA2	QA3	QA4	QA5	QA6	QA7	QA8	QA9
N1	✓	✓	✓	✓	✓	✗	✓	✓	✓
N2	✓	✗	✓	✓	✓	✓	✓	✗	✓
N3	✓	✓	✓	✓	✗	✓	✓	✓	✗
N4	✓	✓	✗	✗	✓	✓	✗	✓	✓
N5	✓	✓	✓	✓	✓	✓	✓	✓	✗
N6	✓	✓	✓	✓	✓	✓	✓	✓	✓
N7	✓	✓	✓	✗	✓	✓	✗	✓	✗
N8	✓	✗	✓	✓	✗	✓	✓	✗	✓
N9	✓	✓	✓	✓	✓	✓	✓	✓	✓
N10	✓	✗	✓	✓	✗	✓	✓	✗	✗
N11	✓	✗	✓	✓	✓	✓	✓	✗	✓
N12	✓	✓	✓	✓	✓	✓	✓	✓	✗
N13	✓	✓	✓	✓	✓	✓	✓	✓	✓
N14	✗	✓	✓	✓	✓	✓	✓	✓	✗
N15	✓	✗	✓	✓	✗	✓	✓	✗	✗
N16	✓	✓	✓	✓	✗	✓	✓	✓	✓
N17	✓	✓	✓	✓	✓	✓	✓	✓	✓
N18	✗	✓	✗	✓	✓	✓	✗	✓	✓
N19	✓	✓	✓	✓	✗	✓	✓	✓	✗
N20	✓	✓	✓	✗	✓	✓	✓	✓	✗
N21	✓	✓	✓	✓	✓	✓	✓	✓	✓
N22	✓	✓	✓	✓	✓	✓	✓	✓	✓
N23	✓	✓	✓	✓	✗	✓	✓	✓	✓
N24	✓	✓	✓	✓	✓	✓	✓	✓	✓
N25	✓	✓	✓	✗	✓	✓	✓	✓	✗
N26	✓	✓	✓	✓	✓	✗	✓	✓	✗
N27	✗	✓	✓	✓	✓	✓	✓	✓	✗
N28	✓	✓	✓	✓	✗	✓	✓	✓	✓
N29	✓	✓	✓	✓	✓	✓	✓	✓	✓
N30	✓	✗	✓	✗	✓	✓	✓	✗	✓
N31	✗	✗	✓	✓	✓	✗	✓	✗	✗
N32	✓	✓	✓	✓	✓	✓	✓	✓	✓
N33	✓	✓	✓	✓	✓	✓	✓	✓	✗
N34	✓	✓	✓	✓	✓	✗	✓	✓	✗
N35	✓	✗	✓	✓	✓	✓	✓	✗	✗

Table 5. Quality Assessment

Paper ID	QA1	QA2	QA3	QA4	QA5	QA6	QA7	QA8	QA9
N36	✓	✓	✗	✓	✗	✓	✗	✓	✗
N37	✓	✓	✓	✓	✓	✗	✓	✓	✓
N38	✓	✓	✓	✓	✓	✓	✓	✓	✓
N39	✓	✓	✓	✓	✓	✓	✓	✓	✓
N40	✓	✓	✓	✓	✓	✗	✓	✓	✗
N41	✓	✓	✓	✓	✗	✓	✓	✓	✓
N42	✓	✓	✓	✓	✓	✓	✓	✓	✓
N43	✓	✓	✓	✓	✓	✓	✓	✓	✓
N44	✓	✓	✓	✓	✓	✓	✓	✓	✓
N45	✓	✓	✓	✓	✗	✓	✓	✓	✗
N46	✓	✓	✓	✓	✓	✓	✓	✓	✓
N47	✓	✓	✓	✓	✓	✓	✓	✓	✗
N48	✓	✓	✓	✓	✗	✗	✓	✓	✓
N49	✓	✓	✓	✓	✓	✓	✓	✓	✓
N50	✓	✓	✓	✓	✓	✓	✓	✓	✗
N51	✓	✓	✓	✓	✗	✗	✓	✓	✓
N52	✓	✓	✓	✓	✓	✓	✓	✓	✓
N53	✓	✓	✓	✓	✓	✓	✓	✓	✗
N54	✓	✓	✓	✓	✓	✓	✓	✓	✓
N55	✓	✓	✓	✓	✓	✓	✓	✓	✓
AN1	✓	✓	✓	✓	✓	✗	✓	✓	✗
AN2	✓	✓	✓	✓	✓	✓	✓	✓	✓
AN3	✓	✓	✓	✓	✓	✓	✓	✓	✓
AN4	✓	✓	✓	✓	✓	✓	✓	✓	✓
AN5	✓	✓	✓	✓	✓	✓	✓	✓	✗
AN6	✓	✓	✓	✓	✓	✗	✓	✓	✓
AN7	✓	✓	✓	✓	✗	✓	✓	✓	✓
AN8	✓	✓	✓	✓	✓	✗	✓	✓	✓
AN9	✓	✓	✓	✓	✓	✓	✓	✓	✗
AN10	✓	✓	✓	✓	✓	✓	✓	✓	✓
AN11	✓	✓	✓	✓	✓	✗	✓	✓	✓
AN12	✓	✓	✓	✓	✗	✓	✓	✓	✗
AN13	✓	✓	✓	✓	✓	✓	✓	✓	✓
AN14	✓	✓	✓	✓	✓	✓	✓	✓	✓
AN15	✓	✓	✓	✓	✓	✓	✓	✓	✓
AN16	✓	✓	✓	✓	✓	✓	✓	✓	✓

2. Mislabeled instances: Instances are labeled with wrong class label. This type of errors is common in situations that different classes have similar symptoms.

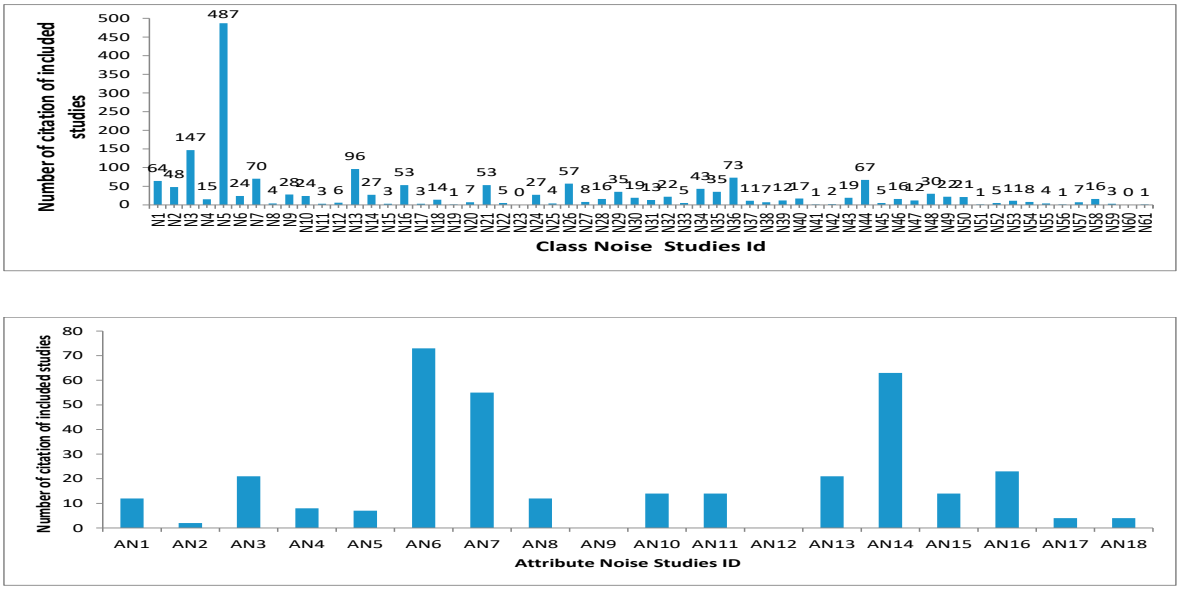

Fig. 2. Citation counts of included studies.

3.3.2. Attribute noise

In contrast to class noise, attribute noise reflects erroneous values for one or more attributes (independent variables) of the data set. Different types of attribute noises are there:

Three types of attribute noise are distinguished: erroneous attribute values, missing or don't know values and incomplete or don't care values[71]. The effectiveness and efficiency of DM/ML methods largely depend on the

Table 6. Class Noise Identification studies.

Class Noise Identification Techniques	Paper ID
Distance based	N12, N13, N14, N15, N16, N17, N18, N47, N50.
Single learning based	N1, N2 ,N3, N4, N6, N8 ,N9, N10, N11, N24, N27, N46, N36, N53.
Ensemble based	N5, N7, N20, N21, N22, N23, N25, N26, N28, N29, N39, N45.

Table 7. Class Noise Handling Studies

Class Noise Handling Techniques	Paper ID
Robust	N30, N31, N38, N40, N51, N52, N53
Filtering	N30, N35, N37, N38, N40, N44, N46
Polishing	N29, N30, N32, N33, N36, N37, N38, N40, N41, N42, N43, N44, N48, N49

quality of the data sets used to build the data mining models. This makes the quality of the data set a significant issue in practice.

3.4. RQ2: Which ML techniques has been used for noise identification?

In this section we have included studies those are belonging to identification techniques of class and attribute noise in data-sets.

3.4.1. RQ2.1: Which techniques are used for identification of class noise?

Class Noise Identification Techniques. There are three main categories of noise identification techniques: Ensemble techniques, distance based algorithm and single learning based techniques to identify noisy instances. Distance-based techniques use closeness measures to ascertain the separation between instances from an information set and utilize this data to distinguish conceivable noisy instances in data. In Ensemble techniques are used to predict noisy instances (mislabeled) in original data sets [7]. In these methods, multiple classifiers are employed to detect mislabeled data. They assume that multiple classifiers tend to generate conflicting class labels for the mislabeled examples. Instead of using distance based and ensemble learning, there are some methods that are based on a single classifier, like a decision tree or neural network.

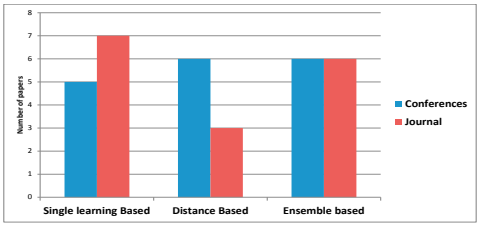

Fig. 3. No. of paper published in Noise Identification.

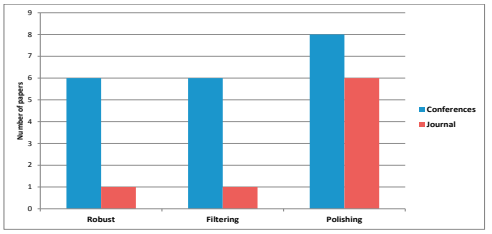

Fig. 4. No. of paper published in Noise Handling.

Now we discuss each techniques which are evolved in the studies individually.

Attribute Noise Identification Techniques. A very few methods are available for detecting instances with attribute noise, largely due to the high complexity of the problem. There are different techniques to identified attribute noise proposed by the [71]. In pairwise attribute algorithm (PANDA) yields a relative ranking of instances from the most to least noisy. PANDA is attractive because it can be used with or without knowledge of class labels, in contrast to other noise detection procedures [47].

4. RQ3: Which ML techniques has been used for handling noise?

After the identification of noise our next task is to handle these noise. In this section we discuss about the different handling techniques.

Handling Class Noise. After the identification of noisy instances, our next task is to handle these noisy instances. In this section, we discuss the different handling techniques. There are three techniques to handle noise in data sets: Noise can be ignored, whereas the techniques analysis have to be robust enough to cope with over-fitting. Noise can be filtered out of the data set after its identification, or it can be altered. The last approach is also called polishing or data scrubbing or relabeling.

Handling Attribute Noise. After the identification of noise in attribute next we use these techniques to handle these attributes. In this section we discuss different techniques which help us to handle missing, erroneous and irrelevant or redundant attribute values. In general, methods to handle erroneous attribute values belong either to filtering or polishing. Filtering is removing erroneous attribute values from data set. Imputation is to identify noise for an attribute, imputation predicts what the clean value is and identify other values suspicious.

5. Discussion

In comparison with class noise, the attribute noise is usually less harmful, but could still bring severe problems to data analysis. Robust algorithms are common in the machine learning community, where the standard approach to coping with imperfections is to delegate the burden to the theory builder. The data set retains the noisy instances, and each algorithm must institute its own noise-handling routine to ensure robustness, duplicating the effort required even if using the same data set in each case. Polishing can help, but when performed improperly, it can also introduce phantom features into the data. In filtering out the noisy instances, there is an obvious trade off between the amount of noise removed and the amount of data retained.

In the extreme case, where every instance is in some way less than perfect, the whole data set might be discarded, leaving us with nothing to analyze. Identification of class noise by distance based techniques mostly based on nearest neighbor algorithm; easy to understand and implement but for example, assume that the examples that are close to each other tend to have the same label; this assumption is not valid for all the data sets. The accuracy of ensemble techniques to detect mislabeled data is usually better than the other two methods as multiple learning algorithms can complement with each other but it consuming more time because multiple classifiers need to be trained. Single learning based method in terms of efficiency, usually better because it is more suitable for the highly dynamic noisy data.

6. Conclusions

The issue of dealing with noise finds applications in numerous domains, where it is attractive to determine interesting and irregular occasions in the movement which produces such information.

The basic techniques are discussed in this paper have also been searched extensively, and have been studied widely in the literature. Details of these methods provided in this paper to identify and handle noise in data sets. Identification and handling noise is important for researchers and practitioners to accurately handle their data and predict future trends. The issue of handle noise turns out to be particularly testing, when huge connections exist among the distinctive information focuses. Noisy data investigation has tremendous scope for research, especially in the area of structural and temporal analysis.

References

[1] Kitchenham, Barbara. (2004) "Procedures for performing systematic reviews." Keele, UK, Keele University **33(2004)**: 1-26.

[2] Zhang, Peng, et al. (2011) "Robust ensemble learning for mining noisy data streams." *Decision Support Systems*, 50(2): 469-479.

[3] Gamberger, Dragan, Nada Lavrac, and Saso Dzeroski. (2000) "Noise detection and elimination in data preprocessing: experiments in medical domains." *Applied Artificial Intelligence*, 14(2): 205-223.

[4] Rebbapragada, Umaa, and Carla E. Brodley. (20007) "Class noise mitigation through instance weighting." *European Conference on Machine Learning. Springer Berlin Heidelberg.*

[5] Zhu, Xingquan, Xindong Wu, and Qijun Chen. (2003) "Eliminating class noise in large datasets." *ICML*, **3**.

[6] Zeng, Xinchuan, and Tony Martinez. (2003) "A noise filtering method using neural networks." *Soft Computing Techniques in Instrumentation, Measurement and Related Applications, 2003. SCIMA 2003. IEEE International Workshop on. IEEE.*

[7] Verbaeten, Sofie, and Anneleen Van Assche. (2003) "Ensemble methods for noise elimination in classification problems." *International Workshop on Multiple Classifier Systems. Springer Berlin Heidelberg.*

[8] Sluban, Borut, Dragan Gamberger, and Nada Lavra. (2010) "Advances in class noise detection." *Proceedings of the 2010 conference on ECAI 2010: 19th European Conference on Artificial Intelligence. IOS Press.*

[9] Zhu, Xingquan, Xindong Wu, and Qijun Chen. (2006) "Bridging local and global data cleansing: Identifying class noise in large, distributed data datasets." *Data mining and Knowledgediscovery*, **12(2-3)**: 275-308.

[10] Zhu, Xingquan, and Xindong Wu. (2006) "Class noise handling for effective cost-sensitive learning by cost-guided iterative classification filtering." *IEEE Transactions on Knowledge and Data Engineering,* **18(10)**: 1435-1440.

[11] Libralon, Giampaolo Luiz, Andre Carlos Ponce de Leon Carvalho, and Ana Carolina Lorena. (2009) "Pre-processing for noise detection in gene expression classification data." *Journal of the Brazilian Computer Society,* **15(1)**: 3-11.

[12] Segata, Nicola, et al. (2010) "Noise reduction for instance-based learning with a local maximal margin approach." *Journal of Intelligent Information Systems*, **35(2)**: 301-331.

[13] Libralon, Giampaolo L., Andre C. Ponce Leon Ferreira Carvalho, and Ana C. Lorena. (2008) "Ensembles of pre-processing techniques for noise detection in gene expression data." *International Conference on Neural Information Processing. Springer Berlin Heidelberg.*

[14] Kim, Sunghun, et al. (2011) "Dealing with noise in defect prediction." *Software Engineering (ICSE), 2011 33rd International Conference on. IEEE.*

[15] Moosavi, M. R., et al. (2010) "An adaptive nearest neighbor classifier for noisy environments." *Electrical Engineering (ICEE), 2010 18th Iranian Conference on. IEEE.*

[16] Liu, Huawen, and Shichao Zhang. (2012) "Noisy data elimination using mutual k-nearest neighbor for classification mining." *Journal of Systems and Software,* **85(5)**: 1067-1074.

[17] Bernadt, Joseph, and Leen Kiat Soh. (2004) "Authoritative citation knn learning with noisy training datasets." *Proceedings of the International Conference on Artificial Intelligence, IC-AI'04*.

[18] Miranda, Andre LB, et al. (2009) "Use of classification algorithms in noise detection and elimination." *International Conference on Hybrid Artificial Intelligence Systems. Springer Berlin Heidelberg.*

[19] Melville, Prem, et al. (2004) "Experiments on ensembles with missing and noisy data." *International Workshop on Multiple Classifier Systems. Springer Berlin Heidelberg.*

[20] Sluban, Borut, Dragan Gamberger, and Nada Lavrac. (2014) "Ensemble-based noise detection: noise ranking and visual performance evaluation." *Data Mining and Knowledge Discovery,* **28(2)**: 265-303.

[21] Garcia, Luis Paulo F., Ana Carolina Lorena, and Andre CPLF Carvalho. (2012) "A study on class noise detection and elimination." *Neural Networks (SBRN), 2012 Brazilian Symposium on. IEEE.*

[22] Smith, Michael R., and Tony Martinez. (2011) "Improving classification accuracy by identifying and removing instances that should be misclassified." *Neural Networks (IJCNN), The 2011 International Joint Conference on. IEEE.*

[23] Zhong, Shi, Wei Tang, and Taghi M. Khoshgoftaar. (2005) *Boosted noise filters for identifying mislabeled data.* Technical report, Department of computer science and engineering, Florida Atlantic University.

[24] Khoshgoftaar, Taghi M., Jason Van Hulse, and Amri Napolitano. (2011) "Comparing boosting and bagging techniques with noisy and imbalanced data." *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans,* **41(3)**: 552-568.

[25] Jeatrakul, Piyasak, Kok Wai Wong, and Chun Che Fung. (2010) "Data cleaning for classification using misclassification analysis." *Journal of Advanced Computational Intelligence and Intelligent Informatics,* **14(3)**: 297-302.

[26] Guan, Donghai, et al. (2011) "Identifying mislabeled training data with the aid of unlabeled data." *Applied Intelligence,* **35(3)**: 345-358.

[27] Teng, Choh-Man. (2001) "A Comparison of Noise Handling Techniques." *FLAIRS Conference.*

[28] Folleco, Andres, et al. (2008) "Identifying learners robust to low quality data." *Information Reuse and Integration, 2008. IRI 2008. IEEE International Conference on. IEEE.*

[29] Teng, Choh Man. (2004) "Polishing blemishes: Issues in data correction." *IEEE Intelligent Systems,* **19(2)**: 34-39.

[30] Sun, Jiang-wen, et al. (2007) "Identifying and correcting mislabeled training instances." *Future generation communication and networking (FGCN 2007),* **1**.

[31] Khoshgoftaar, Taghi M., Shi Zhong, and Vedang Joshi. (2005) "Enhancing software quality estimation using ensemble-classifier based noise filtering." *Intelligent Data Analysis,* **9(1)**: 3-27.

[32] Kubica, Jeremy, and Andrew W. Moore. (2003) "Probabilistic noise identification and data cleaning." *ICDM.*

[33] Teng, Choh-Man. (2003) "Applying noise handling techniques to genomic data: A case study." *Data Mining, 2003. ICDM 2003. Third IEEE International Conference on. IEEE.*

[34] Zhang, Yan, et al. (2005) "ACE: an aggressive classifier ensemble with error detection, correction and cleansing." *Tools with Artificial Intelligence, 2005. ICTAI 05. 17th IEEE International Conference on. IEEE.*

[35] Liebchen, Gernot, Bhekisipho Twala, and Martin Shepperd. (2007) "Filtering, robust filtering, polishing: Techniques for addressing quality in software data." *Empirical Software Engineering and Measurement, 2007. ESEM 2007. First International Symposium on. IEEE.*

[36] Shah, Shital, and Andrew Kusiak. (2010) "Relabeling algorithm for retrieval of noisy instances and improving prediction quality." *Computers in biology and medicine,* **40(3)**: 288-299.

[37] Teng, Choh. (2003) "Noise correction in genomic data." *Intelligent Data Engineering and Automated Learning*, 60-67.

[38] Zeng, Xinchuan, and Tony R. Martinez. (2001) "An algorithm for correcting mislabeled data." *Intelligent data analysis,* **5(6)**: 491-502.

[39] Muhlenbach, Fabrice, Stephane Lallich, and Djamel A. Zighed. (2004) "Identifying and handling mislabelled instances." *Journal of Intelligent Information Systems,* **22(1)**: 89-109.

[40] Sluban, Borut, Dragan Gamberger, and Nada Lavrac. (2010) "Performance analysis of class noise detection algorithms." *Proceedings of the 2010 conference on STAIRS 2010: Proceedings of the Fifth Starting AI Researchers' Symposium. IOS Press.*

[41] Cao, Jingjing, Sam Kwong, and Ran Wang. (2012) "A noise-detection based AdaBoost algorithm for mislabeled data." *Pattern Recognition,* **45(12)**: 4451-4465.

[42] Ipeirotis, Panagiotis G., et al. (2014) "Repeated labeling using multiple noisy labelers." *Data Mining and Knowledge Discovery,* **28(2)**: 402-441.

[43] Teng, Choh. (2000) "Evaluating noise correction." *PRICAI 2000 Topics in Artificial Intelligence,* 188-198.

[44] Lallich, Stephane, Fabrice Muhlenbach, and Djamel A. Zighed. (2002) "Improving classification by removing or relabeling mislabeled in-

stances." *International Symposium on Methodologies for Intelligent Systems. Springer Berlin Heidelberg.*
[45] Smith, Michael R., and Tony Martinez. (2014) "Becoming More Robust to Label Noise with Classifier Diversity." *arXiv preprint arXiv:1403.1893*
[46] Wang, Dong, and Xiaoyang Tan. (2014) "Robust Distance Metric Learning in the Presence of Label Noise." *AAAI.*
[47] Fefilatyev, Sergiy, et al. (2012) "Label-noise reduction with support vector machines." *Pattern Recognition (ICPR), 2012 21st International Conference on. IEEE.*
[48] Saez, Jose A., et al. (2016) "INFFC: an iterative class noise filter based on the fusion of classifiers with noise sensitivity control." *Information Fusion,* **27**: 19-32.
[49] Saez, Jose A., Julian Luengo, and Francisco Herrera. (2016) "Evaluating the classifier behavior with noisy data considering performance and robustness: the Equalized Loss of Accuracy measure." *Neurocomputing,* **176**: 26-35.
[50] Garcia, Luis PF, Andre CPLF de Carvalho, and Ana C. Lorena. (2016) "Noise detection in the meta-learning level." *Neurocomputing,* **176**: 14-25.
[51] Garcia, Luis PF, Andre CPLF de Carvalho, and Ana C. Lorena. (2015) "Effect of label noise in the complexity of classification problems." *Neurocomputing,* **160**: 108-119.
[52] Cesa-Bianchi, Nicolo, Shai Shalev-Shwartz, and Ohad Shamir. (2010) "Online learning of noisy data with kernels." *arXiv preprint arXiv:1005.2296.*
[53] Sluban, Borut, and Nada Lavrac. (2015) "Relating ensemble diversity and performance: a study in class noise detection." Neurocomputing **160**: 120-131.
[54] Sabzevari, Maryam, Gonzalo Martinez-Munoz, and Alberto Suarez. (2015) "Small margin ensembles can be robust to class-label noise." *Neurocomputing,* **160**: 18-33.
[55] Khoshgoftaar, Taghi M., and Jason Van Hulse. (2005) "Identifying noise in an attribute of interest." *Machine Learning and Applications, 2005. Proceedings. Fourth International Conference on. IEEE.*
[56] Sun, Ling, Jia-yu Chi, and Zhong-fei Li. (2006) "A study on reduction of attributes based on variable precision rough set and information entropy." *Machine Learning and Cybernetics, 2006 International Conference on. IEEE.*
[57] Mannino, Michael, Yanjuan Yang, and Young Ryu. (2009) "Classification algorithm sensitivity to training data with non representative attribute noise." *Decision Support Systems,* **46(3)**: 743-751.
[58] Li, Jiye, and Nick Cercone. (2006) "Comparisons on different approaches to assign missing attribute values." *Technical Report CS-2006-04, School of Computer Science, University of Waterloo.*
[59] Khoshgoftaar, Taghi M., and Jason Van Hulse. (2005) "Identifying noisy features with the pairwise attribute noise detection algorithm." *Intelligent Data Analysis,* **9(6)**: 589-602.
[60] Ali, Kamal M., and Michael J. Pazzani. (1993) "HYDRA: A noise-tolerant relational concept learning algorithm." *IJCAI.*
[61] Van Hulse, Jason D., Taghi M. Khoshgoftaar, and Haiying Huang. (2007) "The pairwise attribute noise detection algorithm." K*nowledge and Information Systems,* **11(2)**: 171-190.
[62] Wah, Catherine, and Serge Belongie. (2013) "Attribute-based detection of unfamiliar classes with humans in the loop." Proceedings of the IEEE Co*nference on Computer Vision and Pattern Recognition.*
[63] Khoshgoftaar, Taghi M., Naeem Seliya, and Kehan Gao. (2004) "Rule-based noise detection for software measurement data." , Proceedings of the 2004 IEEE *International Conference on Information Reuse and Integration, IEEE.*
[64] Almasi, Omid Naghash, and Modjtaba Rouhani. (2016) "A new fuzzy membership assignment and model selection approach based on dynamic class centers for fuzzy SVM family using the firefly algorithm." Turkish Journal of Electrical Engi*neering and Computer Sciences,* **24(3)**: 1797-1814.
[65] Mannino, Michael, Yanjuan Yang, and Young Ryu. (2009) "Classification algorithm sensitivity to training data with non representative attribute noise." *Decision Support Systems,* **46(3)**: 743-751.
[66] Goldman, Sally A., and Robert H. Sloan. (1995) "Can pac learning algorithms tolerate random attribute noise?." *Algorithmica,* **14(1)**: 70-84.
[67] Khoshgoftaar, Taghi M., and Jason Van Hulse. (2009) "Empirical case studies in attribute noise detection." *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews),* **39(4)**: 379-388.
[68] Yang, Ying, Xindong Wu, and Xingquan Zhu. (2004) "Dealing with predictive-but-unpredictable attributes in noisy data sources." *European Conference on Principles of Data Mining and Knowledge Discovery. Springer Berlin Heidelberg.*
[69] Lukaszewski, Tomasz, et al. (2011) "Handling the description noise using an attribute value ontology." *Control and Cybernetics,* **40**: 275-292.
[70] Yin, Hua, Hongbin Dong, and Yuxuan Li. (2009) "A cluster-based noise detection algorithm." *Database Technology and Applications, 2009 First International Workshop on. IEEE.*
[71] Zhu, Xingquan, and Xindong Wu. (2004) "Class noise vs. attribute noise: A quantitative study." *Artificial intelligence review,* **22(3)**: 177-210.
[72] Lorena, Ana C., and Andre CPLF de Carvalho. (2004) "Evaluation of noise reduction techniques in the splice junction recognition problem." *Genetics and Molecular Biology,* **27(4)**: 665-672.
[73] Pelletier, Charlotte et al. (2017) A Effect of Training Class Label Noise on Classification Performances for Land Cover Mapping with Satellite Image Time Series. *Remote Sensing,* **9**: 173.
[74] Yuan, Weiwei, Donghai Guan, Tinghuai Ma, and Asad Masood Khattak. (2018) "Classification with class noises through probabilistic sampling." *Information Fusion,* **41**: 57-67.