ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 4, April 2019

Improving Label Noise Filtering by Exploiting Unlabeled Data using Supervised Learning

Mokshada Kotwal[1], Shraddha Khonde[2]

ME Student, Department of Computer Engineering, MES College of Engineering, Pune, Maharashtra, India[1]

Asst. Professor, Department of Computer Engineering, MES College of Engineering, Pune, Maharashtra, India[2]

ABSTRACT: As more and more applications produce streaming data, clustering data streams has become an important technique for data and knowledge engineering. An increasing amount of training data is available in many machine learning tasks. However, it is difficult to ensure perfect labeling with a large volume of training data. Some labels can be incorrect, resulting in label noise, which could lead to deterioration in learning performance. A way to address label noise is to apply noise filtering techniques to identify and remove noise prior to learning. Multiple noise filtering approaches have been proposed. However, the existence of unlabeled data is ignored because most work did on mislabeled data. In fact, unlabeled data are common in many applications, and their values have been extensively studied and recognized. Therefore, in this work, we explore the effective use of unlabeled data to improve the noise filtering performance. To this end, we propose a novel noise filtering algorithm called reinforcement Learning base different Learning algorithms, which is an ensemble-learning-based filter that adopts a soft majority voting strategy. Proposed Learning provides a systematic way to measure the value of unlabeled data by considering different aspects, such as label confidence and the sample distribution. Finally, the effectiveness of the proposed method is confirmed by experiments and comparison with other methods.

KEYWORDS: Label noise; Unlabeled; Majority Voting

I. INTRODUCTION

The ideal situation for extracting knowledge from a dataset is when it does has neither outliers nor noise. However, real-world databases are usually dirty and a cleaning process is required before performing a data mining task. The noise in a dataset may deteriorate the performance of a classifier applied on it, since the misclassification error as well as the computing time may increase and the classifier and decision rules obtained could be more complex. In a supervised classification context, the quality of a dataset is characterized by two information sources: the predictor attributes and the categorical attribute which defines the classes. The quality of the predictors is determined by their quality to represent the instances to be classified, and the quality of the class attribute is determined by the correct assignment of each instance. The quality of a dataset is determined by internal and external factors. The internal factor reveals if the predictors and the classes has been correctly selected and are well defined. The external factor measures errors introduced in the predictors or in the class assignment, either systematically or artificially. In particular, an instance contains noise when it causes problems due to external reasons.

The text mining studies are gaining more importance recently because of the availability of the increasing number of the electronic documents from a variety of sources. The resources of unstructured and semi structured information include the word wide web, governmental electronic repositories, news articles, biological databases, chat rooms, digital libraries, online forums, electronic mail and blog repositories. Therefore, proper classification and knowledge discovery from these resources is an important area for research. The noise contained in the training dataset can be divided into two main categories: attribute noise and label noise. Attribute noise is defined as an imprecision or a mistake introduced in the attribute values, while label noise is caused by mislabeling. The two types of noise have been comprehensively studied in many works, which have suggested that removing attribute noise can decrease the predictive accuracy of a classifier if the same attribute noise is present when the classifier is subsequently used.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 4, April 2019

However, eliminating label noise consistently improves the predictive accuracy. This work focuses on label noise only, and ``noise'' in this work uniquely refers to label noise. This work we will show the effect of different kinds of noise on text classification performance by doing detailed experiments on synthetic as well as real life noisy datasets. Here we are essentially reporting our observations based on experiments and not proposing any new method to combat noise in text for text classification.

II. RELATED WORK

Geetha, A and N. Rajkumar [1] proposed A Statistical Clustering Data Streams Based On Shared Density among Micro Clusters.Micro-clusters formulates shared density enhance by providing the information the knowledge the information of huge data points during an outlined place. On the prevailing demand, a (enhanced) supposed to convey agglomeration algorithmic rule that is employed during a specific offline step to create the micro-clusters into immense final clusters. to create agglomeration, the information of the small clusters are used as pseudo points with clusters isn't keep within the on-line method and re agglomeration is predicated on specific engorged assumptions concerning the promotion of information among and between small clusters that incontestable captures the density between small clusters via information streams supported shared density graph.Carnein, Matthias et.al. [2] proposed An empi rical comparison of stream clustering algorithms."Analyzing streaming data has received considerable attention over the recent years. A key research area in this field is stream clustering which aims to recognize patterns in a possibly unbounded data stream of varying speed and structure. Over the past decades a multitude of new stream clustering algorithms have been proposed According to Mansalis, Stratos, et al. [3] An evaluation of data stream clustering algorithms." Data stream clustering is a hot research area due to the abundance of data streams collected nowadays and the need for understanding and acting upon such sort of data. Unsupervised learning (clustering) comprises one of the most popular data mining tasks for gaining insights into the data. Clustering is a challenging task, while clustering over data streams involves additional challenges such as the single pass constraint over the raw data and the need for fast response. Moreover, dealing with an infinite and fast changing data stream implies that the clustering model extracted upon such sort of data is also subject to evolution over time. Hahsler, Michael, Matthew Bolanos, and John Forrest. [4] proposed Introduction to stream: An Extensible Framework for Data Stream Clustering Research with R." a research tool that includes modeling and simulating data streams as well as an extensible framework for implementing, interfacing and experimenting with algorithms for various data stream mining tasks. The main advantage of stream is that it seamlessly integrates with the large existing infrastructure provided by R. In addition to data handling, plotting and easy scripting capabilities, R also provides many existing algorithms and enables users to interface code written in many programming languages popular among data mining researchers (e.g., C/C++, Java and Python).Kumar K. Naveen et. Al [5] proposed "Appraise on Various Clustering Modules of Clustering Data Streams based on Shared Density betw2`een Micro-Clusters". This paper describes DBSTREAM, the first micro-cluster-based online clustering component that explicitly captures the density between micro-clusters via a shared density graph. The density information in this graph is then exploited for reclustering based on actual density between adjacent micro-clusters. Using shared density improves clustering quality over other popular data stream clustering methods which require the creation of a larger number of smaller micro-clusters to achieve comparable results.According to Hahsler, Michael, and Matthew Bolanos[6] "Clustering data streams based on shared density between micro-clusters" A typical approach is to summarize the data stream in real-time with an online process into a large number of so called micro-clusters.Chitty, Avula. "Efficiency of Clustering Data Streams Based on Micro-Clusters Shared Density" in [7]. A normal approach is to summarize the data stream in real-time with an online process into countless called micro-clusters. Micro-clusters represent local density estimates by assemble the information of many data points in a defined area. On request, a (modified) traditional clustering algorithm is used in a second offline step to re-cluster the micro clusters into larger final clusters. For re-clustering, the coordinator of the micro-clusters is used as pseudo points with the density estimates used as their weights.According to Carnein, Matthias, and Heike Trautmann. "evoStream–Evolutionary Stream Clustering Utilizing Idle Times." Big Data Research [8]. Clustering possibly unbounded and evolving data streams is of particular interest due to the widespread deployment of large and fast data sources such as sensors. The vast majority of stream clustering algorithms employ a two-phase approach where the stream is first summarized in an online phase. Upon request, an offline phase re-clusters the aggregations into the final clusters. In this setup, the online component

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 4, April 2019

will idle and wait for the next observation in times where the stream is slow. Fahy et. Al. [9] "Ant colony stream clustering: A fast density clustering algorithm for dynamic data streams. A stream is potentially unbounded, data points arrive online and each data point can be examined only once. This imposes limitations on available memory and processing time. Furthermore, streams can be noisy and the number of clusters in the data and their statistical properties can change over time. This paper presents an online, bio-inspired approach to clustering dynamic data streams. The proposed ant colony stream clustering (ACSC) algorithm is a density-based clustering algorithm, whereby clusters are identified as high-density areas of the feature space separated by low-density areas. ACSC identifies clusters as groups of micro-clusters. According to Tari, Zahir, et al. [10] "MicroGRID: An Accurate and Efficient Real-Time Stream Data Clustering with Noise." Data stream clustering aims to produce clusters from a data-stream in a real-time. Many of existing algorithms focus however on solving a single problem, leaving anomalous noise in data streams at the wayside. This paper describes the MicroGRID approach to cluster data from single data-streams to handle noisy data streams, accurately identifying and separating noise-affected data points from outlier points. In particular, MicroGRID utilizes a combination of micro-cluster and grid-based prospective, an approach that has not been attempted when clustering data-streams.

III. DISCUSSION

Basically we study the two type of noise in our proposed survey after the training phase has successfully done.
a) Noise in the attributes: It is given by the errors occurred during the entrance of the values of the attributes. Among the sources of this type of noise are: variables with missing values, and redundant data.
b) Noise in the classes: It is given by the errors introduced during the assignment of the instances to the classes. The presence of this kind of noise may be due to subjectivity, errors in the data entry process, and incorrect information for assigning a instance to a class. There are two possible sources of class noise: contradictory examples and misclassifications.
i) Inconsistent instances. These are instances with the same attribute values but belonging to two or more different classes of the dataset, and ii) Error in the classification. Instances assigned incorrectly to a class. This type of error usually happens when there are classes with similar values for the attributes.
The proposed algorithm identifies the noisy instances and distinguishes them from the instances that are in class boundary. The goal of the algorithm is to identify and delete the noisy instances, preserving the class distribution and the classes boundaries such that the neither separability of the classes nor the discriminate power of the classification algorithm is altered. The presence of noise in the classes affects significantly the performance of a classifier, since it modifies the class boundaries and it becomes more difficult to determine them. The instances representing noise cause that a classifier assigns incorrect classes to instances that are correctly labeled. Thus, noisy instance have a direct effect on the accuracy of the classifier.

IV. CONCLUSION AND FUTURE WORK

This paper shows that different noise removing methods have much to offer, especially in terms of scalability to large datasets. It exemplify this with a novel chaining method that can model label correlations while maintaining acceptable computational complexity. Empirical evaluation over a broad range of multi-label datasets with a variety of evaluation metrics demonstrates the competitiveness of our chaining method against related and state-of-the-art methods, both in terms of predictive performance and time complexity. Below things we consider for future enhancement.

- Noise insertion in the most influence parameter during the classification.
- System proposed the noise filter approach for training phase.
- Automated text categorization application.
- Purpose of Automatic Text Classification to learn machine to classify a text.
- Map the text to one or more predefined categories .

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 4, April 2019

- Similar to extraction semantics of the text.
- Learn a classifier from a set of labeled bags.

REFERENCES

1. Geetha, A., and N. Rajkumar. "A Statistical Clustering Data Streams Based On Shared Density among Micro Clusters." International Journal of Research 5.6 (2018): 258-263.
2. Carnein, Matthias, Dennis Assenmacher, and Heike Trautmann. "An empirical comparison of stream clustering algorithms." Proceedings of the Computing Frontiers Conference. ACM, 2017..
3. Mansalis, Stratos, et al. "An evaluation of data stream clustering algorithms." Statistical Analysis and Data Mining: The ASA Data Science Journal 11.4 (2018): 167-187.
4. Hahsler, Michael, Matthew Bolanos, and John Forrest. "Introduction to stream: An Extensible Framework for Data Stream Clustering Research with R." Journal of Statistical Software 76.14 (2017): 1-50.
5. KUMAR, K. NAVEEN, Mr RAGHU KUMAR LINGAMALLU, and G. SREENIVAS. "Appraise on Various Clustering Modules of Clustering Data Streams based on Shared Density between Micro-Clusters"." (2018).
6. Hahsler, Michael, and Matthew Bolaños. "Clustering data streams based on shared density between micro-clusters." IEEE Transactions on Knowledge and Data Engineering 28.6 (2016): 1449-1461.
7. Chitty, Avula. "Efficiency of Clustering Data Streams Based on Micro-Clusters Shared Density." (2017)
8. Carnein, Matthias, and Heike Trautmann. "evoStream–Evolutionary Stream Clustering Utilizing Idle Times." Big Data Research (2018).
9.] Fahy, Conor, Shengxiang Yang, and Mario Gongora. "Ant colony stream clustering: A fast density clustering algorithm for dynamic data streams." IEEE Transactions on Cybernetics (2018).
10. Tari, Zahir, et al. "MicroGRID: An Accurate and Efficient Real-Time Stream Data Clustering with Noise." Pacific-Asia Conference on Knowledge Discovery and Data Mining. Springer, Cham, 2018.