International Journal of Research in Advent Technology, Special Issue, March 2019
E-ISSN: 2321-9637
International Conference on Technological Emerging Challenges (ICTEC-2019)
Available online at www.ijrat.org

Anoptimistic Probabilistic Learning Approach For Class Imbalance Data Sources

Mr. Gummadidala Jaya Krishna[1] , Mrs. Kona Krishna Priya[2], Ms. Turaka Padmasri[3]
[1,2,3]Computer science and Engineering
[1]IIIT Srikakulam, RGUKT-APSrikakulam, India
[1,2]IIIT Nuzvid -RGUKT-AP Nuzvid, India
jayakrishna@rguktsklm.ac.in, indiakrishnapriyavaliveti06@gmail.com, padmasri.turaka@gmail.com

Abstract: In this paper, an optimistic approach for class imbalance learning scenario is presented. The approach is known as Optimistic Probabilistic Learning (OPL) for Class Imbalance, which uses the unique technique for finding the optimal threshold values for reducing the effect of class imbalance on the probabilistic approach of naïve bayes. The specific techniques of optimal threshold havesuccesses in many scenarios due to the data specific intrinsic details evaluation and formulation for optimistic learning. The experimental results are conducted on the six varied datasets and the results suggest that an improved performance can be achieved using the proposed method.

Keywords—Data Mining, Classification, Naïve Bayes, Skewed Data, Optimal Probabilistic Learning.

1. INTRODUCTION

Data mining is the field of knowledge discovery from the existing databases. The knowledge discovered from the vast data sources can be efficiently utilized for many novel reasons. The varieties of data sources for mining knowledge are increasing day by day. The unique data sources, such as class imbalance are difficult for processing in finding the new knowledge.

Indeed all the papers on data mining consider mining from a data source which is having almost equal instances in all the classes. This stereo- type scenario of data representation exists only in the artificial data generation or in the data representation where exact numbers of instances are placed in all classes manually. In the real time scenario, there is no chance of having equal number of instance in the classes.

The rest of this paper is organized as follows: Section 2 presents the concept of class imbalance learning. Section 3 presents the main related work about Bayesian classifier. Section 4 provides a detailed explanation of the Optimistic Probabilistic Learning (OPL) for Class Imbalance.Section 5 presents the datasets used for experiments. Section 6 presents the experimental results. Section 7 draws the conclusions and points out future research.

2. THE BACKGROUND

One of the most popular techniques for alleviating the problems associated with class imbalance is data sampling. Data sampling alters the distribution of the training data to achieve a more balanced training data set. This can be accomplished in one of two ways: under sampling or oversampling. Under sampling removes majority class examples from the training data, while oversampling adds examples to the minority class. Both techniques can be performed either randomly or intelligently.

The random sampling techniques either duplicate (oversampling) or remove (under sampling) random examples from the training data. Synthetic minority oversampling technique (SMOTE) [1] is a more intelligent oversampling technique that creates new minority class examples, rather than duplicating existing ones. Wilson's editing (WE) intelligently undersamples data by only removing examples that are thought to be noisy. In this study, we investigate the impact of unique under and oversampling technique on the performance of the classification algorithms. While the impacts of noise and imbalance have been frequently investigated in isolation, their combined impacts havenot received enough attention in research, particularly with respect to classification algorithms. To alleviate this deficiency, we present a comprehensive empirical investigation of learning from noisy and imbalanced data using classification techniques.

Finding minority class examples effectively and accurately without losing overall performance is the objective of class imbalance learning. The fundamental issue to be resolved is that the classification ability of most standard learning algorithms is significantly compromised by imbalanced class distributions. They often give high overall accuracy, but form very specific rules and exhibit poor generalization for the within class. Correspondingly, the majority class is often over generalized. Particular attention is necessary for each class. It is important to know if a performance improvement happens to both classes and just one class alone.

3. RELATED WORK

Rosa Blanco *et al.* [2] have proposed a filter and wrapper approaches based on the feature subset selection are adapted to induce Bayesian classifiers (naïve Bayes, selective naive Bayes, semi naive Bayes, tree augmented naive Bayes, and k-dependence Bayesian classifier) and are applied to distinguish between the two subgroups of cirrhotic patients. Heni Bouhamed et al. [3] have proposed ansolution whereby a remedy can be conceived for the intricate algorithmic complexityimposed during the learning of Bayesian classifiers structure with the use of sophisticated algorithms.

Christophe Salperwyck et al. [4] have proposed a new method basedon a graphical model which computes the weights on the input variables using astochastic estimation. The method is incremental and produces an Weighted Naïve Bayes Classifier for data stream. This method will be compared to classical naïve Bayes classifier on the Large Scale Learning challenge datasets.Karl-Michael Schneider et al. [5] have discussed Naive Bayes as often used in text classification applications and experiments because of its simplicity and effectiveness. However, its performance is often degraded because it does not model text well, and by inappropriate feature selection and the lack of reliable confidence scores. They address these problems and show that they can be solved by some simple corrections.

Wei Zhang et al [6] have analyzed the performance of naïve bayes in text classification and the corresponding results from different points of view are proposed, then an improving way for text classification with highly asymmetric misclassification costs is provided.WeiZhang et al [7] have proposed an auxiliary feature method which determines features by an existing feature selection method, and selects an auxiliary feature which can reclassify the text space aimed at the chosen features. Sona Taheri et al., [8] have proposed an algorithm which approximates the interactions between features by using conditional probabilities.J. N. K. Liu et al., [9] have proposed an improved naive Bayes classifier (INCB) technique and explores the use of genetic algorithms (GAs) for the selection of a subset of input features in classification problems.

4. FRAMEWORK OF OPTIMISTIC LEARNING FOR CLASS IMBALANCE SCENARIO

The following are the different stages for Optimistic Probabilistic Learning (OPL) for Class Imbalance approach.

A Naïve Bayesian classifier uses the technique of learning probabilistic for decision tree building. The Naïve Bayesian classifier is one of the specialized approaches for supervised learning task where the aim is to correctly classify the unseen instances. The classifier follows these two simple assumptions for accurate prediction. First, the predicting attributes are conditionally independent for a given class. Second, is that no hidden or latent attributes influence the predicting process.

If there are C1, C2 classes and some A1, A2, A3… attributes then Naïve Bayesian classifier uses the simple probabilistic rule to compute the probability of each class for a specific predictive attribute for classification. The technique works well for general datasets where the numbers of instances in the classes are same.If the datasets is in the class imbalance scenario, then Naïve Bayesian classifier performance decreases. To boost the performance, a specific optimistic learning technique is to be implemented where the intrinsic nature of data set are to be upgraded.

In the initial stage the data source is partitioned into different sub group of instances probably binary in nature. The sub group which has more percentage of instances can be termed as majority subset and the sub group with less percentage of instances is known as minority subset. The noisy or missing values instances are to be removed from both the sub groups as they help for improvement of the quality of the dataset.

In the later stage, majority set can be considered for under sampling strategy, which is used to remove the excessive instances from the subset. The idea of performing under sampling is an effective, in terms of reduction of high percentage of instances from the majority subset. One more prominent strategy of performing over sampling in minority subset also solves the problem of class imbalance to some extent. The over sampling technique generates new instances in the minority subset to reduce the imbalance ratio of the overall dataset.

The above said techniques have provided solid evidences for improving the probabilistic approach of the naïve bayes classifier. The same techniques are incorporated in the existing naïve byes classifier for our new proposal Optimistic Probabilistic learning (OPL).

5. DATASETS

In the study, we have considered 7 binary data-sets which have been collected from the UCI [10] machine learning repository Web sites, and they are very varied in their degree of complexity, number of attributes, number of instances, and imbalance ratio (the ratio of the size of the majority class to the size of the minority class). The number of attributes ranges from 9to 29, the number of instances ranges from 57 to 3772, and the imbalance ratio is from 1.70 to 11.58. This way, we have different IRs: from low imbalance to highly imbalanced data-sets. Table 1 summarizes the properties of the selected data-sets: for each dataset, S.no, Dataset name, the number of examples (#Ex.), number of attributes (#Atts.), class name of each class (minority and majority) and the IR. This table is ordered according to the name of the datasets in alphabetical order.

Table 1 Summary of benchmark imbalanced datasets

S.no	Datasets	# Ex.	# Atts.	Class (_,+)	IR
1.	wisconsin-breast-cancer	699	9	(bening; maligant)	1.90
2.	horse-colic.ORIG	368	27	(1 ; 2)	1.96
3.	horse-colic	368	22	(yes ; no)	1.70
4.	hungarian-14-heart	294	13	(positive ; negative)	1.70
5.	hepatitis	155	19	(die; live)	3.84
6.	labor	57	16	(positive ; negative)	1.85
7.	sick	3772	29	(positive ; negative)	11.58

We have obtained the accuracy and other metric estimates by means of a 10-fold cross-validation. That is, the data-set was split into ten folds, each one containing 10% of the patterns of the dataset. For each fold, the algorithm is trained with the examples contained in the remaining 9 folds and then tested with the current fold. The data partitions used in this paper can be found in UCI-dataset repository [10] so that any interested researcher can reproduce the experimental study.

6. EXPERMENTAL RESULTS

The experimental results of the proposed approach OPL and the traditional naïve bayes classifier are presented in this section. The results of the both methods are generated in equal terms of data source and experimental setup. This equal terms simulation will help to investigate the limitations and strengths of the proposed approach OPL on varied data sources. Table 2 presents the results of OPL in comparison with Naïve Bayes in terms of accuracy. The performance of OPL method is improved for all the datasets, suggest that optimistic probabilistic learning solves the issue of decrease in accuracy of classification for naïve bayes. The result details of AUC, root mean square error are presented in table 3 and 4 and the performance is improved.

Table 2 Summary of tenfold cross validation performance for accuracy on all the datasets

Dataset	Naïve Bayes	OPL
wisconsin-breast-cancer	96.12	**97.30**
horse-colic.ORIG	72.33	**76.77**
horse-colic	77.39	**79.27**
hungarian-14-heart	82.83	**83.44**
hepatitis	83.29	81.94
labor	93.87	**95.07**
sick	92.89	**94.90**

Table 3 Summary of tenfold cross validation performance for AUC on all the datasets

Dataset	Naïve Bayes	OPL
wisconsin-breast-cancer	0.988	**0.993**
horse-colic.ORIG	0.829	0.819
horse-colic	0.838	**0.849**
hungarian-14-heart	0.906	0.894
hepatitis	0.847	**0.873**
labor	0.984	0.980
sick	0.930	**0.951**

Table 4 Summary of tenfold cross validation performance for Root mean Square error on all the datasets

Dataset	Naïve Bayes	OPL
wisconsin-breast-cancer	0.185	**0.145**
horse-colic.ORIG	0.446	**0.410**
horse-colic	0.418	**0.404**
hungarian-14-heart	0.227	**0.225**
hepatitis	0.351	0.354
labor	0.142	**0.183**
sick	0.226	0.640

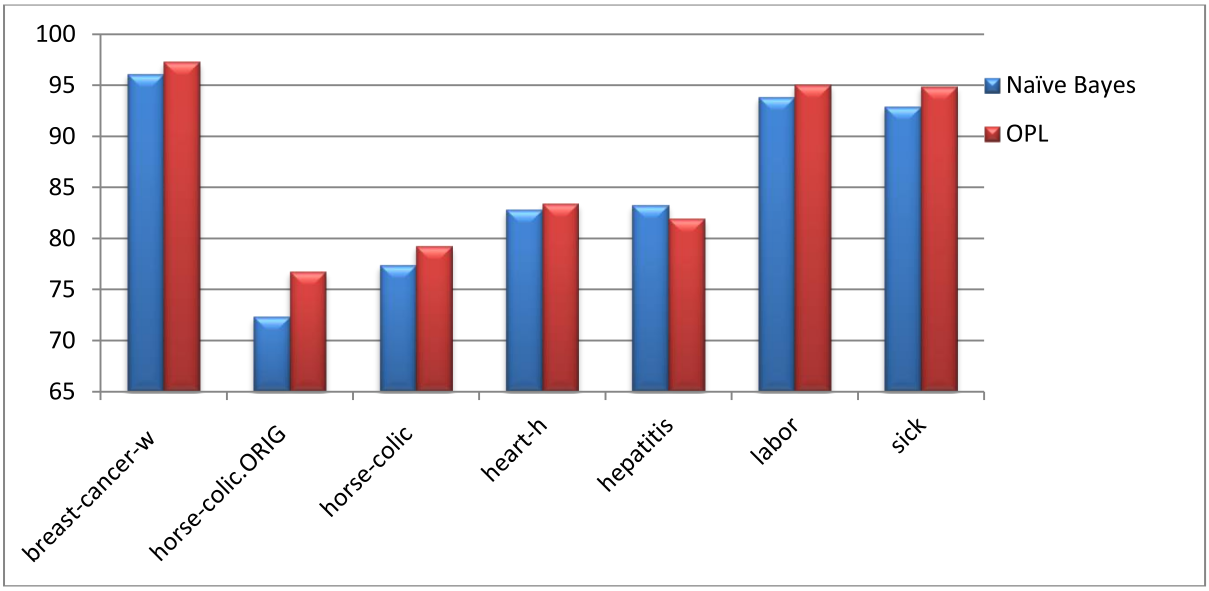

Fig. 1 Test results on accuracy between the Naïve Bayes verses OPL on all the datasets.

The results of accuracy are summarized in the figure 1. From figure 1, one can understand that the performance of the proposed approach is improved than the existing naïve bayes algorithm. The improved in the performance suggest that an optimal value of probabilistic learning can be one of the best solutions for better applicability of naïve bayes for class imbalance data sources.

7. CONCLUSION

In this paper, a novel approach for imbalanced distributed data has been proposed. This method uses unique optimistic threshold learning to reduce the effect of class imbalance scenario. Empirical results have shown that the proposed OPL considerably reduces the non uniform effect of the datasets while retaining or improving the performance measure when compared with benchmark method.

REFERENCES

[1] N. Chawla, K. Bowyer, and P. Kegelmeyer, “SMOTE: Synthetic minority over-sampling technique,” *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002.

[2] Rosa Blanco, In˜aki Inza, Marisa Merino , Jorge Quiroga , Pedro Larran˜aga,” Feature selection in Bayesian classifiers for the prognosisof survival of cirrhotic patients treated with TIPS”, Journal of Biomedical Informatics 38 (2005) 376–388.

[3] Heni Bouhameda, Afif Masmoudib, Ahmed Rebai,” Bayesian classifier structure-learning using several general Algorithms”, Procedia Computer Science 46 (2015) 476 – 482.

[4] Christophe Salperwyck1, Vincent Lemaire2, and Carine Hue,” Incremental Weighted Naive Bayes Classiffer for Data Stream”,

[5] Karl-Michael Schneider,” Techniques for Improving the Performance of Naive Bayes for Text Classification”,

[6] Wei Zhang and Feng Gao, "Performance analysis and improvement of naïve Bayes in text classification application," *IEEE Conference Anthology*, China, 2013, pp. 1-4.doi:10.1109/ANTHOLOGY.2013.6784818.

[7] Wei Zhang, Feng Gao,,”An Improvement to Naive Bayes for Text Classification,”, Procedia Engineering,Volume 15, 2011, Pages 2160-2164, ISSN 1877-

7058,https://doi.org/10.1016/j.proeng.2011.08.404.

[8] Taheri, Sona & Mammadov, Musa & Bagirov, A.M.. (2011). Improving Naive Bayes classifier using conditional probabilities. 9th Australasian Data Mining Conference. 121. 63-68.

[9] J. N. K. Liu, B. N. L. Li and T. S. Dillon, "An improved naive Bayesian classifier technique coupled with a novel input solution method [rainfall prediction]," in *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 31, no. 2, pp. 249-256, May 2001.doi: 10.1109/5326.941848.

[10] Blake C, Merz CJ (2000) UCI repository of machine learning databases. Machine-readable data repository. Department of Information and Computer Science, University of California at Irvine, Irvine. http://www.ics.uci.edu/mlearn/MLRepository.html