Improving the Accuracy of Classifier with the Ensemble of Centroid Based Algorithms

Phyo Thu Thu Khine, Htwe Pa Pa Win
University of Computer Studies, Yangon
phyothuthukhine@gmail.com, hppwucsy@gmail.com

Abstract

Classification accuracy improvement is an essential process in data analysis problems and large amount of data generated in extra bytes per year are produced by many real world application. There are many enhancement techniques to address the problems as regarding the classification performance, have been proposed previously by many researchers. However, the issues of mislabeling problems that depend on noise dataset, class imbalanced problems and big data problems still have been a challenge of today's research. In data mining process, the ensemble of classifiers are known to increase the accuracy of single classifiers by combining several of them, but neither of these learning techniques alone solve the above problems. Therefore, this paper analyzes on the problems of data and emphasizes on the ensemble techniques by using centroid based clustering methods. In order to get the more accuracy, filtering process is used by enhancing the simple ensemble ways of classifier.

1. Introduction

Classification is an important technique used in data mining and data analysis applications. It is used when a set of data has to be separated into predefined classes based on various attributes of the data. In practical applications, classification has been successfully applied to industrial, business and scientific problems such as bankruptcy prediction, credit scoring, medical diagnosis, quality control, handwritten character recognition, and speech recognition. Other applications also included recommendation systems which used classification techniques to provide the most suitable content to users [1].

Today, many real world applications are deal with large amount of data and the collected data included with many noise. Although many Machine Learning (ML) algorithms can deal with noise, detecting and removing noisy instances from the training dataset can help the induction of the target hypothesis [2].

Big Data generated in extra bytes per year has become a watchword of today's research. They are exceptionally afar from the capability of commonly used software tools and also beyond the handling possibility of the single machine architecture. Facing this challenge has activated a requisite to reexamine the data by helping to discover large unknown values from enormous datasets. Also, outmoded management systems and statistical packages express trouble handling Big Data.

In numerous real applications, Handling of imbalanced datasets is the fact of precedence. The classification of datasets having imbalanced class distribution has produced a notable drawback in performance obtained by the most standard classifier learning algorithms. Assuming balanced class distribution and equal misclassification costs lead to poor results. In a real-world domain, the classification methods of multi-class imbalance problem need more attention compared to the two-class problem [3].

The classification task can be carried out by various techniques. The choice of the best technique to a specific problem can be decided by experimenting many possibilities based on the measures such as accuracy, speed, robustness, scalability and interpretability [4]. Using ensemble methods is one of the general strategies to improve the accuracy of classifier and predictor. Ensemble learning is a simple, useful and effective meta-classification methodology that combines the predictions from multiple base classifiers (or learners) [5]. As the advantages of Centroid based clustering method have been established by many researchers, this paper adopted these on ensemble way in order to get the required accuracy improvements.

This paper is organized as followed. The literature reviews are presented in section 2. The problem statements existing on datasets are described

in Section 3. Section 4 describes the background necessary in this proposed system. The proposed methodology is presented in section 5. The experimental results are discussed in section 6. Finally, section 7 ends with conclusions.

2. Literature Reviews

Various methods have been proposed for improving the accuracy of the classification available in the literature and they are discussed in this section.

The group in [6] made the survey of data mining techniques for improvement of prediction accuracy. They described the detail about various classification and clustering methods and stated that the prediction accuracy is depending on the technique used in the application.

Ensemble learning is a machine learning process to get better prediction performance by strategically combining the predictions from multiple learning algorithms. Ensembles are known to reduce the risk of selecting the wrong model by aggregating all the candidate models. In the process of improving ensemble accuracy and stability, different techniques have been established. These techniques vary in their approach to treat the training data, the type of algorithms used, and the combination methods followed [7].

The previous researches in [4, 5, 7-12] are proposed to improve the classification accuracy by using ensemble techniques in different ways. The papers [3] and [4] are based on class imbalanced problems and big data. The rest research works are emphasizing on the hybrid methods on data mining techniques. The others [13, 14] are based on the misclassification errors and then to the assembling process. The researchers in [15] proposed a multi attribute depthness similarity estimation technique to classify the data points towards the number of classes to overcome this more false classification issue.

Because of their accuracy-oriented design, ensemble learning algorithms that are directly applied to imbalanced data-sets do not solve the problem that underlay in the base classifier by themselves. However, their combination with other techniques to tackle the class imbalance problem has led to several proposals in the literature, with positive results. The above big data problem is proposed in [3, 16]. The work in [10] used Feature selection method to improve the classification accuracy.

3. Problem Statements

The performance of the classification accuracy is decreased in the presence of by the following problems.

3.1. Problem of Noise in Datasets

Noise can be defined as an example apparently inconsistent with the remaining examples in a dataset. The presence of noise in a dataset can decrease the predictive performance of Machine Learning (ML) algorithms, by increasing the model complexity and the time necessary for its induction. Datasets with noisy instances are common in real world problems, where the data collection process can produce noisy data [2].

3.2. Problem of Imbalanced Datasets

In classification, a dataset is said to be imbalanced when the number of instances which represents one class is smaller than the ones from other classes. Furthermore, the class with the lowest number of instances is usually the class of interest from the point of view of the learning task. This problem is of great interest because it turns up in many real-world classification problems, such as remote-sensing, pollution detection, risk management, fraud detection, and especially medical diagnosis [4].

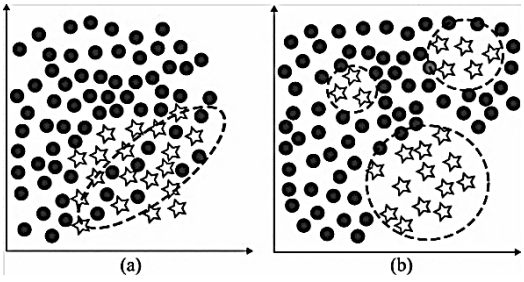

Figure 1. Example of difficulties in imbalanced datasets (a) Class overlapping (b) Small disjuncts

3.3. Problem of Big Data

Big Data is currently defined using three data characteristics: volume, variety and velocity. It means that some point in time, when the volume, variety and velocity of the data are increased, the current techniques and technologies may not be able to handle storage and processing of the data. At that point the data is defined as Big Data. Many applications suffer from the Big Data problem,

including network traffic risk analysis, geospatial classification and business forecasting [16].

3.4. Problem of Classification

In machine learning, multiclass or multinomial classification is the problem of classifying instances into one of three or more classes. (Classifying instances into one of the two classes is called binary classification). While some classification algorithms naturally permit the use of more than two classes, others are by nature binary algorithms; these can, however, be turned into multinomial classifiers by a variety of strategies [17].

4. Preliminaries (Theory Background)

4.1. Ensemble Technique

An ensemble for classification is a composite model, made up of a combination of classifiers. The individual classifiers vote and a class label prediction is returned by the ensemble based on the collection of votes. Ensembles tend to be more accurate than their component classifiers. An ensemble combines a series of k learned models (or base classifiers), M1, M2… Mk, with the aim of creating an improved composite classification model, M*. A given dataset, D, is used to create k training sets, D_1, D_2… D_k, where D_i ($1 \leq i \leq k$ -1) is used to generate classifier Mi. Given a new data tuple to classify, the base classifiers each vote by returning a class prediction. The ensemble returns a class prediction based on the votes of the base classifiers [18].

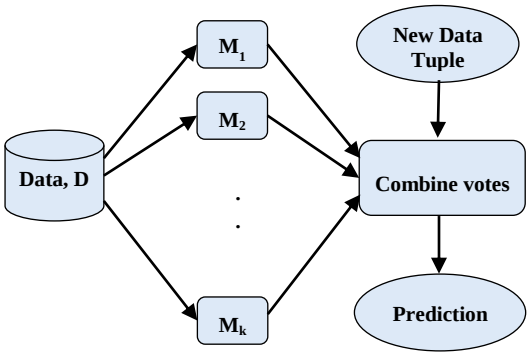

Figure 2. Ensemble Technique

4.2. Classification Algorithms

The classification is the supervised learning algorithms. The state of the art algorithms such as Naive Bayes, Logistic Regression, Decision Tree, Support Vector Machines, and Artificial Neural Network are used for classification model.

4.3. Centroid Based Clustering Algorithm

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group are more similar (in some sense) to each other than to those in other clusters. The clustering is the unsupervised learning algorithms.

In centroid-based clustering, clusters are represented by a central vector, which may not necessarily be a member of the dataset. When the number of clusters is fixed to k, clustering gives a formal definition as an optimization problem: find the k cluster centers and assign the objects to the nearest cluster center, such that the squared distances from the cluster are minimized [17]. Among the centroid based method, this paper used K-Means and EM algorithms to find the clusters.

5. Proposed System

The proposed framework, which focuses on ensemble techniques based on mislabeling analysis in order to improve the classification accuracy, is illustrated in figure 3.

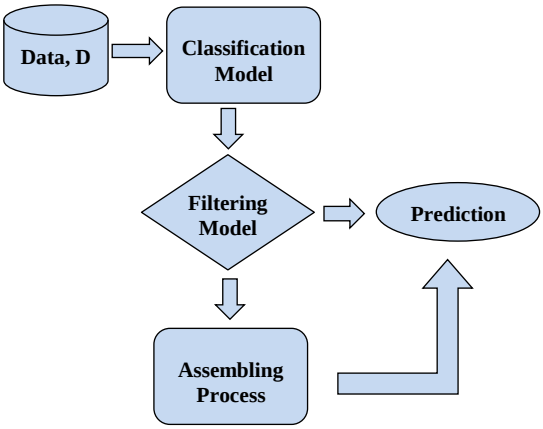

Figure 3. Block diagram of the proposed assemble technique

The proposed system consists of the following processes.

Table 1. Benchmark Dataset Description

Dataset Number	Datasets	Description	Number of Instances	Number of Attributes	Number of Class	Class Distribution
I	Chess (KRKPA7)	Identify White can win or not in chess game	3196	36	2	52% : win, 48% : nowin
II	Mushroom Database	Types of Mushroom	8124	22	2	51.8% : edible, 48.2%: poisonous
III	Pulsars Dataset (HTRU2)	Describes a sample of pulsar candidates	17,898	8	2	9.16%:positive 91.84%:negative
IV	Image Segmentation Data	Identify 26 capital letters in the English alphabet	2310	19	7	Almost equally distributed for all class
V	Thyroid Disease Records	Identify the people have Thyroid disease or not and the level of disease	3772	30	4	92.29%:negative, 5.14%: compensated, 2.52%: primary, 0.05%: secondary
VI	Waveform Database Generator	Identify the classes of wave	5000	21	3	Almost equally distributed for all class

5.1. Dataset Preparation

The real-world datasets used in this paper were obtained from the University of California Irvine (UCI) machine learning repository [19]. These include the datasets for both binary class and multiclass problems and class imbalanced problems, the big dataset problems from different areas as shown in Table 1.

5.2. Classification Model

In the first stage, the classification model is built by using the state of the art algorithms such as Naïve Bayes, Logistic Regression, Decision Tree, Support Vector Machine and Artificial Neural Network, on all the various problem types of dataset.

5.3. Filtering Model

The uncertain, ambiguous, incomplete, and subjective data can reduced the performance of the classifier and not all the techniques are suitable for all type of datasets. Therefore, the misclassification data are filtering out after the classification process. The classification results are got by comparing with the real outputs and testing data. Knowledge from this analysis will be stored in the separate database and used with the testing data. During the testing stage, similarity and matching techniques are applied with misclassified knowledge in order to filter the test data. Then the misclassified data are forwarded to the assembling process.

5.4. Assembling Process

In this process, ensemble classification techniques using majority vote algorithm is applied to for the mislabeling data. Therefore, the centroid based methods are used as the booster to analyses the mislabeling data, the weakness sign of the classifier.

5.5. Prediction

One of the results from the filtering process is directly sent to this process, and another output from the ensemble process is made the prediction at this stage. The prediction for the classifiers is promoted by the powerful clusters' prediction, at the misclassified data.

6. Results and Experiments

After the datasets have been prepared, the next step is to build the classification and clustering model as a preliminary stage for future prediction. There were 5 different types of algorithms implemented for classification, Naïve Bayes (NB), Logistic Regression (Logistic), Decision Tree (DT), Artificial Neural Networks (ANN) and Support Vector Machine (SVM) and 2 different types of centroid based algorithm, K-Means cluster and Expectation Maximization cluster. The performance accuracy of the original classifiers and clusters for the datasets are measured and depicted as in Table 2 and Table 3.

Table 2. Accuracy of Clusters

Dataset Number	K-Means	EM
I	52.23	60.86
II	62.37	89.71
III	91.86	86.74
IV	58.87	61.3
V	49.33	52.25
VI	51.1	51.23

Table 3. Accuracy of Classifiers

Dataset Number	Classification Accuracy %				
	NB	Logistic	DT	SVM	ANN
I	88.32	97.96	99.65	96.24	99.93
II	95.88	100	100	100	100
III	94.52	97.94	98.27	97.60	98.17
IV	80.47	96.70	98.91	93.29	97.53
V	95.44	97.40	99.81	93.69	96.89
VI	80.22	87.64	97.5	87.68	98.06

Based on the results in the above tables, all the type of classifier can handle the problems of datasets. But, DT and ANN outperform the other techniques for both binary and multiclass datasets, including the class imbalanced problems and big data problem. However, the time taken is needed to consider the type of application used because, ANN takes more time to model and DT has a little time consuming. Therefore DT is the best for all the classifiers in this paper.

As the results from Table 2 and 3, it is clear that there is a need to improve the accuracy results and then the data are set to the proposed system model. And the results are measured again as in Table 4. The records are separately displayed according to the centroid method used. The time taken is needed to compare the analysis purpose and depicted in Table 5, the time are in seconds. The times are measured for each dataset separately.

Average performance comparison for each classifier in accuracy is illustrated in figure 4. The accuracy data for classifier alone is known as simple, the ensemble techniques with K-Means is With KM and also with Expectation Maximization is With EM. It can be clearly observed that ensemble technique with EM outperforms than K-Means algorithm but take more time.

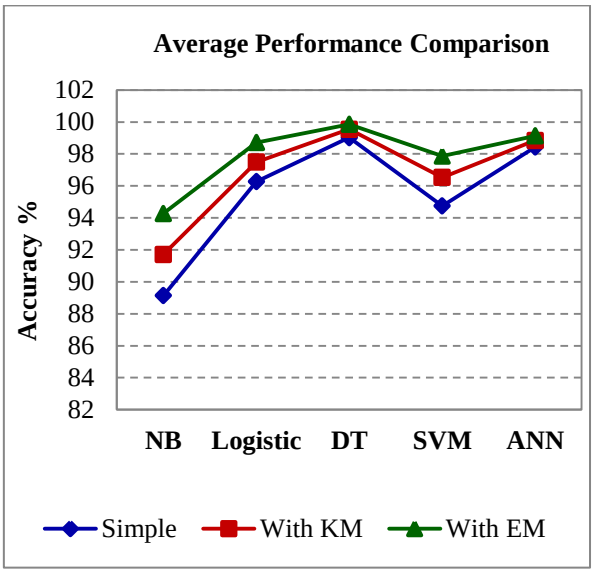

Figure 4. Average accuracy performance comparison among the algorithms

7. Conclusion

Choosing the most suitable classification methods is an important issue in data mining and data analysis applications of the real world. However, construction of the model could be difficult due to different nature of the datasets. Therefore, suitable techniques and methodologies are needed to improve the accuracy of the classification model. In this paper, different problems of dataset are analyzed with the help of UCI repository. The ensemble methods based on centroid clustering is proposed. The results from the comparison of five well known used classification techniques with centroid based methods on each dataset are reported. The performance of each ensemble technique is increased for all typed of datasets and described that the used mechanisms can handle for all nature of dataset problems.

Table 4. Accuracy of Proposed System

Dataset Number	Classification Accuracy %									
	NB+ KMeans	NB+ EM	Logistic+ KMeans	Logistic+ EM	DT + KMeans	DT + EM	SVM+ KMeans	SVM+ EM	ANN+ KMeans	ANN+ EM
I	89.89	91.45	98.84	99.71	100.0	100	99.49	100	99.96	100
II	96.06	96.23	100	100	100	100	100	100	100	100
III	95.35	96.18	98.32	98.70	99.62	100	98.11	98.61	99.45	100
IV	82.94	85.41	97.79	98.87	99.09	99.26	93.76	94.24	97.79	98.051
V	95.97	96.50	98.22	99.04	99.86	99.92	95.46	97.24	97.79	98.70
VI	90.1	99.89	91.8	95.96	98.72	99.94	92.38	97.08	98.06	98.06

Table 5. Time taken Comparison

		Dataset					
Classifier		I	II	III	IV	V	VI
Naïve Bayes (NB)	Simple	0.05	0.06	0.12	0.13	0.09	0.17
	NB + KM	0.09	0.1	0.19	0.22	0.11	0.28
	NB + EM	0.18	0.13	0.28	0.35	0.19	0.89
Logistic Regression (L)	Simple	0.02	0.56	0.6	0.02	0.03	0.07
	L + KM	0.04	0	0.09	0.03	0.04	0.13
	L + EM	0.05	0	0.19	0.06	0.12	0.46
Decision Tree (DT)	Simple	0.01	0.54	0.03	0.01	0.01	0.02
	DT + KM	0.02	0	0.05	0.03	0.02	0.05
	DT + EM	0.02	0	0.12	0.04	0.02	0.1
Support Vector Machines (SVM)	Simple	0.27	0.56	0.35	0.03	0.04	0.11
	SVM + KM	0.04	0	0.05	0.07	0.06	0.17
	SVM + EM	0.09	0	0.21	0.11	0.19	0.51
Artificial Neural Networks (ANN)	Simple	0.03	1.17	0.05	0.03	0.05	0.06
	ANN + KM	0.04	0	0.08	0.04	0.07	0.09
	ANN + EM	0.04	0	0.14	0.07	0.13	0.13

References

[1] W. Paireekreng and K. W.Wong, "Mobile content personalization using intelligent user profile approach", In IEEE Third International Conference on Knowledge Discovery and Data Mining, Phuket, Thailand, January 2010, pp. 241-244, ISBN: 978-0-7695-3923-2.

[2] G.L. Libralon, A.C. Carvalho and A.C. Lorena, "Pre-processing for noise detection in gene expression classification data", Journal of the Brazilian Computer Society, Vol 15, Issue 1, March 2009, pp. 3–11, ISSN 0104-6500.

[3] S.S. Patil and S.P. Sonavane, "Enriched Over_Sampling Techniques for Improving Classification of Imbalanced Big Data", In IEEE Third International Conference on Big Data Computing Service and Applications, San Francisco, USA, April 2017, pp. 65-70, ISBN: 978-1-4577-1479-5.

[4] M. Galar, A. F´andez, E. Barrenechea, H. Bustince and F. Herrera, "A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and

Hybrid-Based Approaches", IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), Volume: 42, Issue: 4, July 2012, pp. 463 – 484, ISSN: 1094-6977.

[5] N. Joshi1 and S. Srivastava, "Improving Classification Accuracy Using Ensemble Learning Technique (Using Different Decision Trees)", International Journal of Computer Science and Mobile Computing (IJCSMC), Volume 3, Issue 5, May 2014, pp. 727–732, ISSN 2320–088X.

[6] P.D. Bagul and Prof. K.C. Waghmare, "A Survey of Data Mining Techniques for Improvement of Prediction Accuracy", International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE), Volume 5, Issue 3, March 2017, ISSN (Print): 2320-9798.

[7] A.O.M. Abuassba, D. Zhang, X. Luo, A. Shaheryar and H. Ali, "Improving Classification Performance through an Advanced Ensemble Based Heterogeneous Extreme Learning Machines", Computational Intelligence and Neuroscience, Hindawi, Volume 2017, DOI:10.1155/2017/3405463.

[8] D. Lavanya and Dr.K.Usha Rani, "A Hybrid Approach to Improve Classification with Cascading of Data Mining Tasks", International Journal of Application or Innovation in Engineering & Management (IJAIEM), Volume 2, Issue 1, January 2013, pp. 345-350, ISSN 2319 – 4847.

[9] P. Pandey and I. Singh, "Improving Accuracy using different Data Mining Algorithms", International Journal of Computer Applications (0975 – 8887), Volume 150, No.10, September 2016, pp. 10-13.

[10] P. Pujari and J. B. Gupta, "Improving Classification Accuracy by Using Feature Selection and Ensemble Model", International Journal of Soft Computing and Engineering (IJSCE), Volume 2, Issue 2, May 2012, ISSN: 2231-2307.

[11] A. Fuxman, A. Kannan, A.B. Goldberg, R. Agrawal, P. Tsaparas and J. Shafer, "Improving Classification Accuracy Using Automatically Extracted Training Data", The 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'09), Paris, France, June 28 - July 01, 2009, ISBN: 978-1-60558-495-9.

[12] A. Ahlawat and B. Suri, "Improving Classification in Data mining using Hybrid algorithm", 1st India International Conference on Information Processing (IICIP), Delhi, India, 12-14 August, 2016, ISBN: 978-1-4673-6985-5.

[13] W. Siriseriwan and K. Sinapiromsaran, "Attributes Scaling for K-Means Algorithm Controlled by Misclassification of all Clusters", Third International Conference on Knowledge Discovery and Data Mining", Phuket, Thailand, 9-10 January, 2010, ISBN: 978-1-4244-5397-9.

[14] L. Rutkowski, M. Jaworski, L. Pietruczuk, and Piotr Duda, "A New Method for Data Stream Mining Based on the Misclassification Error", IEEE Transactions on Neural Networks and Learning Systems, Volume 26, No. 5, May 2015, pp. 1048-1059.

[15] N. Elavarasan ; K. Mani, "An Enhanced Multi Attribute Depthness Similarity Estimation Technique to Improve Classification Accuracy", World Congress on Computing and Communication Technologies (WCCCT), Tiruchirappalli, India, 2-4 Feb, 2017, pp. 115-118, ISBN: 978-1-5090-5574-6.

[16] S. Suthaharan, "Big Data Classification: Problems and Challenges in Network Intrusion Prediction with Machine Learning", ACM SIGMETRICS Performance Evaluation Review, Volume 41, No. 4, March 2014, pp. 70-73, DOI: 10.1145/2627534.2627557.

[17] S. Khan, ML Researcher, Postdoc @U of Toronto, "What-are-the-best-clustering-algorithms-used-in-machine-learning", https://www.quora.com/, Updated October 4, 2017.

[18] J. Han, M. Kamber and J. Pei, "Data Mining Concepts and Techniques", Third Edition, Morgan Kaufmann, USA, ISBN: 978-0-12-381479-1.

[19] UCI Machine Learning Repository, https://archive.ics.uci.edu/ml/datasets.html