A Review of Multi-Class Classification for Imbalanced Data

Mahendra Sahare[1], Hitesh Gupta[2]
Department of Computer Science & Engineering, PCST, Bhopal

Abstract

Prediction and correct voting is critical task in imbalance data multi-class classification. Accuracy and performance of multi-class depends on voting and prediction of new class data. Assigning of new class of imbalance data generate confusion and decrease the accuracy and performance of classifier. Various authors and research modified the multiclass classification approach such as one against one and one against all. In both method OAO and OAA create a unclassified region for data and decrease the performance of classifier such as support vector machine. Some other method such as decision tree classifier, nearest neighbor and probability based classifier also suffered from imbalance data classification. In this paper we discuss various method and approach for multi-class classification for imbalance data.

Keywords

Multi-class classification, SVM, Imbalance data

1. Introduction

In machine learning multiclass classification is a major problem. Each instance in the learning set belongs to a number of set of previously defined labels in multiclass classification. An electronic document can be referred as political and social topic in text categorization [5]. The nearest neighbor rule is an instance based classifier for statistical learning. Each neighbor is considered as an item of evidence for two-valued mass function. The NN rule decreases the noise level present in the training set. Two methods group for multiclass is categorized: problem transformation and problem adaption. The problem transformation groups transform the problem in binary classification problem. While problem adaption group manipulate multiclass data directly adapting some specific algorithm. Support Vector Machine is a tool for machine learning to solve the problem reorganization problem. SVM training is more complex and it slows down the deployment cycle when no of training sample increases because when number of sample training increases SVM requires more memory. Support Vector Machine as application in image retrieval[7]. SVM formulation has been originally developed for binary classification problems and finding the direct formulation for multi-class case is not easy and stills an ongoing research issue. In two ways we can have a multi-class SVM classifier; the first is to consider all data in one optimization formulation, and the second is to simplified multi-class problem to several binary problems. The second solution is a better one and has been considered more than the first approach because binary classifiers are easier to implement and moreover some powerful Algorithms such as Support Vector Machine (SVM) are inherently binary[9]. The complexity of SVM slows down the computation performance. To reduce the complexity due to increase in number of class, the multiclass classifier is simplified into a series of binary classification such as One-Against-One and One-Against-All. In the binary classification there also exists a problem of imbalanced class distribution [12,13]. Using Data Balance algorithm and One against One technique combined this problem is solved. The imbalanced data problem can be an obstacle for inductive learning machine. So there are two approaches called data level approach and algorithm level approach are there to deal with the data misbalancing problem [15]. The data level approach aims to rebalance the class distribution and is applied before a classifier is trained. While algorithm level approach is used to strengthen the exiting classifier to recognize the small class by adjusting the applied algorithm. The rest of paper is organized as follows. In Section 2 Discuss related work of multi-class classification. The Section 3 problem of multi-class classification 4 discuss conclusion of multiclass classification.

2. Related Work

In recent years multi-class classification field for data imbalancing is front scenario for researchers in the field of machine learning. Various authors propose a technique for the improvement of accuracy and prediction of class in multi-class classification. Some works are summarized here in the form of title and their contribution. Salvador Garcia, Jose Ramon Cano, Alberto Fernandez and Francisco Herrera entitled a method of Prototype Selection for Class Imbalance Problems as [1] classification algorithms is said to be unbalanced when one of the classes is represented by a very small number of cases compared to the other classes when a set of input is provided. In such cases, standard classifiers tend to be flooded by the large classes and ignore the small ones. A number of solutions have been proposed at the data and algorithmic levels. At the data level, we found forms of re-sampling such as over-sampling, where replication of examples or generation of new instances is performed; or under-sampling, where elimination of examples is performed. At the algorithmic level, an adjust of the operation of the algorithm is carried out to treat with unbalanced data.

Zhi-Hua Zhou and Xu-Ying Liu entitled study of Training Cost-Sensitive Neural Networks with Methods Addressing the Class Imbalance Problem as [2] the effect of sampling and threshold-moving in training cost-sensitive neural networks. These techniques modify the distribution of the training data such that the costs of the examples are conveyed explicitly by the appearances of the examples. Threshold moving tries to move the output threshold toward inexpensive classes such that examples with higher costs become harder to be misclassified. In classical machine learning or data mining settings, the classifiers usually try to minimize the number of errors they will make in dealing with new data.

Piyasak Jeatrakul, KokWaiWong, and Chun Che Fung entitled an analysis of Data Cleaning for Classification Using Misclassification [3] as in most classification or function approximation problems; the establishing of an accurate prediction model has always been a challenging problem. When constructing a prediction model, it is always difficult to have an exact function or separation that describes the relationship between the input vector, X and target vector, Y. This paper presents the proposed misclassification technique to increase the confidence of cleaning noisy data used for training. In this paper, we focus our study for classification problem using ANN. The CMTNN is applied to detect misclassification patterns. For our proposed technique, the training data is cleaned by eliminating the misclassification patterns discovered by both the Truth NN and Falsity NN. After misclassification patterns are removed from the training set, a neural network classifier is trained by using the cleaned data.

Amal S. Ghanem and Svetha Venkatesh, Geoff West entitled problem in Multi-Class Pattern Classification in Imbalanced Data [4] as the majority of multi-class pattern classification techniques are proposed for learning from balanced datasets. However, in several real-world domains, the datasets have imbalanced data distribution, where some classes of data may have few training examples compared for other classes. Despite the success of these techniques reported in different domains for various types of applications, such as text document classification, and speech recognition, most of these techniques are mainly proposed for learning from relatively balanced training data. In this paper we focused on two main challenges in pattern recognition: the imbalanced class problem and multi-class classification. We reviewed the different strategies proposed to solve these two challenges. Based on this research, we outlined a framework that can handle these challenges simultaneously. Our approach (Multi-IM) is based on a relational technique designed for the binary imbalanced problem (PRMs-IM). Multi-IM extends PRMs-IM to a generalized framework for multi-class classification.

Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall and W. Philip Kegelmeyer entitled a problem of Synthetic Minority Over-sampling Technique [5] as An approach to the construction of classifiers from imbalanced datasets is described. A dataset is imbalanced if the classification categories are not approximately equally represented. A dataset is imbalanced if the classes are not approximately equally represented. Imbalance on the order of 100 to 1 is prevalent in fraud detection and imbalance of up to 100,000 to 1 has been reported in other applications (Provost & Fawcett, 2001). There have been attempts to deal with imbalanced datasets in domains such as fraudulent telephone calls (Fawcett & Provost, 1996), elecommunications management (Ezawa, Singh, & Norton, 1996), text classification (Lewis & Catlett, 1994; Dumais, Platt, Heckerman, & Sahami, 1998; Mladeni´c & Grobelnik, 1999; Lewis & Ringuette, 1994; Cohen, 1995a) and detection of oil spills in satellite images (Kubat, Holte, & Matwin, 1998). The performance of machine learning algorithms is typically evaluated using predictive accuracy. The results show that the SMOTE approach can improve the accuracy of classifiers for a minority class. SMOTE provides a new approach to over-sampling. The combination of SMOTE and under-sampling performs better than plain under-sampling.

Guobin Ou,Yi Lu Murphey entitled an application of Multi-class pattern classification using neural networks [6] as Multi-class pattern classification has many applications including text document classification, speech recognition, object recognition, etc. Multi-class pattern classification using neural networks is not a trivial extension from two-class neural networks. This paper presents a comprehensive and competitive study in multi-class neural learning with focuses on issues including neural network architecture, encoding schemes, training methodology and training time complexity. He has presented an in-depth study on K-class pattern classification using neural networks. Specifically, we discussed two different architectures, systems of multiple neural networks and single neural network systems, and three types of modeling approaches, OAA, OAO, and PAQ.

Jeatrakul, P. and Wong entitled an Enhancing classification performance of multi-class imbalanced data using the OAA-DB algorithm [7] as combining the multi-binary classification technique called One-Against-All (OAA) and a data balancing technique.
In the experiment, the three multi-class imbalanced data sets used were obtained from the University of California Irvine (UCI) machine learning repository. This paper proposed a technique named as the One-Against-All with Data Balancing (OAA-DB) algorithm to solve the multi-class imbalanced problem. It applies the multi-binary classification techniques called the One-Against-All (OAA) approach and the combined data balancing technique. The combined data balancing technique is the integration of the under-sampling technique using Complementary Neural Network (CMTNN) and the over -sampling technique using Synthetic Minority Over -sampling Technique (SMOTE).

Jeatrakul, P., Wong, K.W., Fung, C.C. and Takama entitled a misclassification analysis for the class imbalance problem [8] as the class imbalance issue normally causes the learning algorithm to be dominated by the majority classes and recognize slightly the minority classes. This will indirect affect how human visualize the data. This paper presents the proposed misclassification technique to re-distribute the data in classes to solve the class imbalance problem. This paper uses ANN as the core technique for classification. The CMTNN is applied to detect misclassification patterns. For the proposed technique I, training data is down sized by eliminating only the misclassification patterns discovered by both the Truth NN and Falsity NN. For technique II, the training data is downsized by eliminating all misclassification patterns discovered by the Truth NN and Falsity NN. These two techniques are applied for under-sampling either the majority class or the minority classes.

Sofie Verbaeten and Anneleen Van Assche entitled a Methods for Noise Elimination in Classification Problems [9] as ensemble methods combine a set of classifiers to construct a new classifier that is (often) more accurate than any of its component classifiers. In many applications of machine learning the data to learn from is imperfect. Different kinds of imperfect information exist, and several classifications are given in the literature. He addressed the problem of training sets with mislabeled examples in classification tasks. He proposed a number of filter techniques, based on ensemble methods for identifying and removing noisy examples. He experimentally evaluated these techniques on noise-free ILP data sets which we artificially corrupted with different levels of classification noise. We reported results concerning filter precision, tree size and accuracy.

Jeatrakul, P., Wong, K.W. and Fung entitled a classification of imbalanced data by combining the complementary neural network and SMOTE algorithm [10] as when the distribution of the training data among classes is uneven, the learning algorithm is generally dominated by the feature of the majority classes. The features in the minority classes are normally difficult to be fully recognized. In this paper, a method is proposed to enhance the classification accuracy for the minority classes. The proposed method combines Synthetic Minority Over-sampling Technique (SMOTE) and Complementary Neural Network (CMTNN) to handle the problem of classifying imbalanced data. According to Gu et al. [4], there are two main approaches to deal with imbalanced data sets: data-level approach and algorithm approach. While the data-level approach aims to re-balance the class distribution before a classifier is trained, the algorithm level approach aims to strengthen the existing classifier by adjusting algorithms to recognize the smaller classes. This paper presents the proposed combined techniques to re-distribute the data in classes to solve the class imbalance problem. They are the integration of under sampling techniques using Complementary Neural Network (CMTNN) and the oversampling technique using Synthetic Minority Over-sampling Technique (SMOTE).

Jaree Thongkam , Guandong Xu, Yanchun Zhang and Fuchun Huang entitled a prediction model through improving training space for breast cancer

[11] as medical prognoses need to deal with the application of various methods to historical data in order to predict the survivability of particular patients suffering from a disease using traditional analytical applications such as Kaplan–Meier and Cox-Proportional Hazard, over a particular time period (Borovkova, 2002). However, more recently, due to the increased use of computing automated tools allowing the storage and retrieval of large volumes of medical data to be collected and made available to the medical research community, there has been increasing interest in the development of prediction models using a new method of survival analysis entitled period analysis. In this paper, the OOS approach has been proposed and applied to the tasks of building accurate breast cancer survivability prediction models. This approach is a combination of outlier filtering and over-sampling approaches.

Jeatrakul, P., Wong, K.W. and Fung entitled a data cleaning technique to remove noise, inconsistent data and errors in order to obtain a better and representative data set to develop a reliable prediction model [12] as in most prediction model, unclean data could sometime affect the prediction accuracies of a model. In order to apply Complementary Neural Network CMTNN for data cleaning, Truth NN and Falsity NN are employed to detect the misclassification patterns. This paper presents the proposed misclassification technique to increase the confidence of cleaning noisy data used for training. The CMTNN is applied to detect misclassification patterns. In the experiment, the training data is cleaned by two cleaning techniques. For technique I, the training data is cleaned by eliminating all misclassification patterns discovered by the Truth NN and Falsity NN. For technique II, training data is cleaned by eliminating only the misclassification patterns discovered by both the Truth NN and Falsity NN.

Gustavo E. A. P. A. Batista, Ronaldo C. Prati and Maria Carolina Monard entitled a problem of learning from imbalanced data sets [13] as A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data. There are several aspects that might influence the performance achieved by existing learning systems. It has been reported that one of these aspects is related to class imbalance in which examples in training data belonging to one class heavily outnumber the examples in the other class. In this situation the learning system may have difficulties to learn the concept related to the minority class. Most learning systems usually assume that training sets used for learning are balanced. However, this is not always the case in real world data where one class might be represented by a large number of examples, while the other is represented by only a few. This is known as the class imbalance problem and is often reported as an obstacle to the induction of good classifiers by Machine Learning (ML) algorithms. In this work, he evaluates ten different methods of under and over-sampling to balance the class distribution on training data. Two of these methods, random over-sampling and random under-sampling, are non-heuristic methods that were initially included in this evaluation as baseline methods. The main objective of our research is to compare several balancing methods published in the literature, as well as the three proposed methods, in order to verify whether those methods can effectively deal in practice with the problem of class imbalance. His results show that the over-sampling methods in general, and Smote + Tomek and Smote + ENN (two of the methods proposed in this work) in particular for data sets with few positive (minority) examples, provided very good results in practice. Moreover, Random over-sampling, frequently considered an unprosperous method provided competitive results with the more complex methods.

P. Jeatrakul and K.W. Wong entitled a problem of Comparing the Performance of Different Neural Networks for Binary Classification [14] as classification problem is a decision making task where many researchers have been working on. There are a number of techniques proposed to perform classification. Neural network is one of the artificial intelligent techniques that have many successful examples when applying to this problem. This paper presents a comparison of neural network techniques for binary classification problems. In order to solve the classification problems and prediction, many classification techniques have been proposed. Some of the successful techniques are Artificial Neural Networks (ANN), Support Vector Machines (SVM) and classification trees. This paper presents the comparison of five neural network techniques for binary classification on three benchmarking UCI data sets. Each neural network technique selected for this comparison has different structures and different advantages and disadvantages. While RBFNN, GRNN and PNN have simpler architectures and they can train data faster than BPNN, BPNN is a robust model and it can provide competent results in various problems. In addition, CMTNN has special features based on

BPNN. It uses a pair of opposite BPNN to deal with uncertainty. As can be seen from the results, it can be concluded that CMTNN is suitable for the binary classification problems. The results of the experiment show that CMTNN can provide good results in most cases. This is because of their unique features of using the true and false networks. CMTNN can deal with the uncertainty better than other techniques by using a pair of complementing neural network.

Jose G. Moreno-Torres and Francisco Herrera entitled a Study on Overlapping and Data Fracture in Imbalanced Domains by means of Genetic Programming-based Feature Extraction [15] as classification of imbalanced data is a well-studied topic in data mining. However, there is still a lack of understanding of the factors that make the problem difficult. The classification of imbalanced data is a priority issue in the literature nowadays. We have presented GP-RST, a GP-based feature extractor that employs RST techniques to estimate the fitness of individuals. We have shown GP-RST to be a competitive preprocessing method for highly imbalanced datasets, with the added advantage of providing bi-dimensional representations of the datasets it preprocesses, which are easily interpreted. We have, through the analysis of the visual representations of the preprocessed datasets, observed a data fracture problem between training and test sets, especially in the minority class, that is affecting the classification performance.

3. Problem of Multi-Class Classification

In the process of review we found that some serious problem related to the multi-class classification. These problem are affected the performance and accuracy of multi-class classifier and generate unclassified region. The unclassified region increase, decrease the accuracy and performance of classifier. Some problem are mentioned here[4,6,9,10].

1. Infinite population of data.
2. Feature selection of data
3. Voting of class
4. New class generation.
5. imbalanced data problem
6. Error Correcting Code

4. Conclusion and Future Work

In this paper we present survey of multi-class classification for data imbalancing. Here we discuss some problem related to multi-class classification and also discuss their minimization technique. Some problem such as infinite population of data and feature selection process of classifier is critical problem found in review. Implementation of binary classifier in the form of liner classifier generate such a problem, the first approach relied on extending binary classification problems to handle the multiclass case directly. This included neural networks, decision trees, support vector machines, naive bayes, and k-nearest neighbors. The second approach decomposes the problem into several binary classification tasks. Several methods are used for this decomposition: one versus- all, all-versus-all, erorr-correcting output coding, and generalized coding. The third one relied on arranging the classes in a tree, usually a binary tree, and utilizing a number of binary classifiers at the nodes of the tree till a leaf node is reached. In future we minimized the problem of feature reduction problem and error correcting code for binary classifier.

References

[1] P. Jeatrakul and K.W. Wong “Comparing the Performance of Different Neural Networks for Binary Classification Problems” in Eighth International Symposium on Natural Language Processing, 2009.

[2] Zhou, Zhi-Hua, and Xu-Ying Liu. "Training cost-sensitive neural networks with methods addressing the class imbalance problem." Knowledge and Data Engineering, IEEE Transactions on 18, no. 1 (2006): 63-77.

[3] Piyasak Jeatrakul, KokWaiWong, and Chun Che Fung “Data Cleaning for Classification Using Misclassification Analysis” in Data Cleaning for Classification Using Misclassification Analysis, 2010.

[4] Amal S. Ghanem and Svetha Venkatesh, Geoff West “Multi-Class Pattern Classification in Imbalanced Data” in International Conference on Pattern Recognition, 2010.

[5] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall and W. Philip Kegelmeyer “SMOTE: Synthetic Minority Over-sampling Technique” in Journal of Artificial Intelligence Research 16 (2002) 321–357, 2002.

[6] Guobin Ou,Yi Lu Murphey “Multi-class pattern classification using neural networks” in The Journal of the Pattern Recognition Society, 2007.

[7] Jeatrakul, P., Wong and K.W. “Enhancing classification performance of multi-class imbalanced data using the OAA-DB algorithm” in Annual International Joint Conference on Neural Networks (IJCNN), 2012.

[8] Jeatrakul, P., Wong, K.W., Fung, C.C. and Takama in“Misclassification analysis for the

class imbalance problem" INWorld Automation Congress (WAC), 2010.

[9] Verbaeten, Sofie, and Anneleen Van Assche. "Ensemble methods for noise elimination in classification problems." In Multiple classifier systems, pp. 317-325. Springer Berlin Heidelberg, 2003.

[10] Jeatrakul, P., Wong, K.W. and Fung "Classification of imbalanced data by combining the complementary neural network and SMOTE algorithm" in 17th International Conference on Neural Information Processing (ICONIP), 2010.

[11] Jaree Thongkam *, Guandong Xu, Yanchun Zhang, Fuchun Huang "Toward breast cancer survivability prediction models through improving training space" in Expert Systems with Applications, 2009.

[12] Jeatrakul, P., Wong, K.W. and Fung "Using misclassification analysis for data cleaning" in International Workshop on Advanced Computational Intelligence and Intelligent Informatics (IWACIII), 2009.

[13] Batista, Gustavo EAPA, Ronaldo C. Prati, and Maria Carolina Monard. "A study of the behavior of several methods for balancing machine learning training data." ACM Sigkdd Explorations Newsletter 6, no. 1 (2004): 20-29.

[14] P. Jeatrakul and K.W. Wong "Comparing the Performance of Different Neural Networks for Binary Classification Problems" in Eighth International Symposium on Natural Language Processing, 2009.

[15] Moreno-Torres, Jose G., and Francisco Herrera. "A preliminary study on overlapping and data fracture in imbalanced domains by means of genetic programming-based feature extraction." In Intelligent Systems Design and Applications (ISDA), 2010 10th International Conference on, pp. 501-506. IEEE, 2010.