Survey of Novel Process for Domain Adaptation and Cost Sensitive Online Classification in Data Mining

Ashish. S. Kale[1], S. P. Kosbatwar[2]

[1, 2]Department of Computer Engineering, Smt. Kashibai Navale College of Engineering, Vadgaon Bk, Pune, India

Abstract: ***In the society of machine learning and data mining, online learning and cost-sensitive classification are two topics which are very widely studied. Though those topics are very widely studied but important problem like Cost-Sensitive Online Classification is studied very inadequately. In age of big data, there is an urgent need of developing a technique to mine fast increasing data. There are many methods which are extensively studied for machine learning. Main aim of online learning is learn more forecast model which will make correct forecasts/predictions on the flow of examples which will arrive consecutively. Online learning is beneficial because it has high efficiency and scalability. Online learning has been used for solving online classification tasks. These tasks include a wide range of real-world data mining applications. Many Online learning techniques have been implemented for online classification tasks. For example: Perceptron algorithm, Passive-aggressive (PA) learning. Although it is broadly studied, already implemented techniques are not appropriate for cost-sensitive classification tasks. Misclassification costs need to be focused in cost-sensitive classification tasks. Most of already implemented online learning methods are depended on online classification algorithm.***

Keywords: Cost Sensitive Classification, Online Learning, Online Classification, Online Gradient Descent, Online Anomaly Detection.

1. Introduction

For mining fast increasing data, algorithms are needed which will help machine learning and data mining. These algorithms should be scalable and adequate for mining big data. Number of techniques have been proposed by researchers [11] [13] for studying efficient and scalable machine learning methods. Relaxed online maximum margin algorithm (ROMMA) [12] is technique in which hyper plane is chosen repeatedly to classify already seen examples accurately with maximum margin. By using method of minimization of the length of the weight vector subject to a number of linear constraints, maximum margin hypothesis can be calculated. It works by maintaining comparatively simple relaxation of these constraints that can be sufficiently updated. Mistake bound for ROMMA is same as of Perceptron algorithm. The maximum-margin algorithm satisfies this mistake bound. ROMMA also has equal computational complexity and simplicity as of Perceptron algorithm. Their generalization is much better.

For making accurate forecasts, in which examples will arrive in sequential manner, learning of prediction models is essential in online learning. Online learning is proving useful because of its efficiency and scalability for large scale applications. In number of data mining applications, online learning is used for solving online classification tasks.

Online learning systems have been proposed. Famous among them are Passive-Aggressive (PA) learning, Perceptron calculation and number of lately proposed calculations [14] [17] [18]. In spite of being examined broadly, most existing online learning methods are unacceptable for expense touchy grouping assignments, an issue for information mining which need to address differed misclassification costs [15] [16]. The current web learning procedures possibly may not be compelling enough principally on the grounds that generally existing web learning studies regularly concern the execution of an online grouping calculation.

To address problem of classification, experts specifically from information mining area writing have proposed more compelling measurements, for example, the weighted aggregate of sensitivity and specificity [19] and the weighted misclassification cost [16] [20]. Over the previous decades, considerable research exertions have been given to creating batch classification algorithm to enhance the often suffer poor efficiency and scalability when solving large-scale problems, which are not suitable for online classification applications. Although both expense delicate arrangement and web learning have been considered broadly in information mining and machine learning groups, separately, there were not very many thorough studies on "Expense Sensitive Online Grouping" in both information mining and machine learning writing. To solve online cost-sensitive classification task, cost-sensitive algorithms should be developed. The main challenge here is to develop an effective cost-sensitive online algorithm which will be able to optimize a predefined cost-sensitive measure.

2. Literature Survey

The work is mainly related to three groups of research in data mining and machine learning:

2.1 Cost-sensitive classification
2.2 Online learning
2.3 Anomaly detection

2.1 Cost-sensitive classification

In data mining and machine learning, Cost-sensitive classification has been extensively studied [2] [4] [5] [7]. AdaCost [1] is modified version of AdaBoost. AdaCost is misclassification cost-sensitive boosting method. For updating the training distribution on consecutive boosting rounds, AdaCost makes use of cost of misclassifications. The purpose of AdaCost is to reduce misclassification cost than that of AdaBoost. It has observed that AdaCost minimizes the upper bound of misclassification cost. Without using extra computing power, Empirical evaluations have shown significant reduction in the misclassification cost.

Study [3] has shown that the performance of cost-sensitive classifiers gets affected by class-imbalance. Cost-sensitive classifiers support normal class distribution when costs do not differ significantly. On the other hand when cost differs significantly, cost-sensitive classifiers support balanced class distribution. This research tells that when we are handling relatively balanced data set, we can apply cost-sensitive learning, while handling significantly imbalanced data set, we should balance the class distribution.

Rescaling the classes is eminent approach of cost-sensitive learning. Rescaling the classes is obtained in such way that the effects of different classes on the learning process are proportional to their weights. Typical approach contains assigning of training examples of different classes with different weights. In this weights are proportional to the misclassification costs. It has been observed that traditional rescaling approach is ineffective on multi-class approach. It also observed that instead of rescaling directly, the consistency of the costs must be examined [6].

There are many real time cost-sensitive problems present, like medical diagnosis and fraud detection. In the situations like these problems, cost of misclassifying a target is much higher than that of a false positive. Classifiers which are finest under symmetric costs have a tendency to underperform. Researchers have proposed number of cost-sensitive metrics to deal with this problem. Most of popular examples of cost-sensitive metrics include Weighted sum of sensitivity and specificity [19] and also the weighted misclassification cost [16] [20]. While measuring classification performance, Weighted misclassification cost also include cost into consideration. The weighted sum of sensitivity and specificity is reduced to the well-known balanced accuracy [19], when the weights are both equal to 0.5. Balanced accuracy is extensively used in anomaly detection tasks.

2.2 Online learning

A sequence of data examples with time stamps are generally used for Online Learning. Very little work has been done on optimizing the two cost-sensitive metrics in online learning. Except [8] which is based on online Naive Bayes approach. In this work, rare events detection is termed as imbalanced classification problem. It attempts to find events which have high effect but less probability of occurrence. Network intrusion detection and credit fraud detection are applications of rare events detection. They have proposed online algorithm which differs from traditional accuracy oriented methods. To obtain the cost/benefit analysis, this approach uses number of hypothesis tests. This method can deal with online data with unbounded data volumes. For that, they have set a proper moving-window site and a forgetting factor.

Recent online learning study [9], it has been analyzed that traditional online learning methods fail when data is unevenly distributed between different classes, because traditional online learning methods measure performance of a learner by classification accuracy. This limitation is tackled by developing algorithm for maximizing Area Under ROC curve (AUC). AUC is metric which is most extensively used to measure the classification performance for imbalanced data distributions. Main problem in online AUC maximization is that it is necessary to optimize pair by pair loss between two instances from different classes. This is exact opposite of traditional online learning in which total loss is addition of all losses.

2.3 Anomaly Detection

Anomaly detection is also known as outlier detection or novelty detection. The Aim of anomaly detection is to discover unusual data patterns which do not relate to normal patterns. Anomaly detection has been studied widely from last few years. In past works [10], novelty detection in semi-supervised setting is automatically solved by reducing to a binary classification problem. A detector which has desired false positive rate can be accomplished by reduction in to Neyman-Pearson classification. In contrast of inductive method, semi-supervised novelty detection (SSND) defers detectors that are optimal despite of the distribution on novelties. In novelty detection, there is a substantial impact on the theoretical properties of the decision rule of unlabeled data.

2.4 Methodology for Cost-sensitive online classification

We explore the proposed application in depth to tackle emerging the big data mining challenges in domain adaptation. We proposed an algorithm for tackling the problem of domain adaptation. The common problem occurred in online active learning is that the training data and the operational (testing) data are drawn from different underlying distributions. This has a more complexity for many statistical learning methods. Therefore we studied domain adaptation method that parameterizes this concept space by linear transformation under which we explicitly minimize the distribution difference between the source domain with sufficient labeled data and target domains with a large amount of unlabeled data, while simultaneously reducing the empirical loss on the labeled data in the source domain.

3. Mathematical Model

For performance metrics, sensitivity is defined as the ratio between the number of true positives Tp−Mp and the number of positive examples; specificity is defined as the ratio between Tn−Mn and the number of negative examples; and accuracy is defined as the ratio between the number of correctly classified examples and the total number of examples. These can be summarized as:

Sensitivity = (Tp -Mp)/Tp
Specificity = (Tn -Mn)/Tn
Accuracy = (T-M)/T

M = denote the number of mistakes
Mp = denote the number of false negatives,
Mn = denote the number of false positives
T = to denote the set of indexes of negative examples,
Tp = denote the number of positive examples,
Tn = denote the number of negative examples.

4. Conclusion

In this survey paper, we have focused some important issues of cost sensitive online classification. Very limited study addresses cost sensitive online classification. Based on such report, we surveyed Cost-sensitive classification in data mining, Online learning in machine learning, Anomaly detection in both data mining and machine learning. We have analyzed their cost-sensitive bounds. We further examined empirical performance, and studied applications to tackle real-world online anomaly detection tasks.

References

[1] W. Fan, S. J. Stolfo, J. Zhang, and P. K. Chan, "Adacost: Misclassification cost-sensitive boosting," in Proc. 16th ICML, New York, NY, USA, 1999, pp. 97–105.
[2] C. X. Ling, V. S. Sheng, and Q. Yang, "Test strategies for cost sensitive decision trees," IEEE Trans. Knowl. Data Eng., vol. 18, no. 8, pp. 1055–1067, Aug. 2006.
[3] X.-Y. Liu and Z.-H. Zhou, "The influence of class imbalance on cost-sensitive learning: An empirical study," in Proc. 6th ICDM, Washington, DC, USA, 2006, pp. 970–974.
[4] H.-Y. Lo, J.-C. Wang, H.-M. Wang, and S.-D. Lin, "Cost-sensitive multi-label learning for audio tag annotation and retrieval," IEEE Trans. Multimedia, vol. 13, no. 3, pp. 518–529, Jun. 2011.
[5] B. Zadrozny, J. Langford, and N. Abe, "Cost-sensitive learning by cost-proportionate example weighting," in Proc. 3rd IEEE ICDM, Washington, DC, USA, 2003, pp. 435–442.
[6] Z.-H. Zhou and X.-Y. Liu, "On multi-class cost-sensitive learning," Comput. Intell., vol. 26, no. 3, pp. 232–257, 2010.
[7] X. Zhu and X. Wu, "Class noise handling for effective cost sensitive learning by cost-guided iterative classification filtering," IEEE Trans. Knowl. Data Eng., vol. 18, no. 10, pp. 1435–1440, Oct. 2006.
[8] J. H. Zhao, X. Li, and Z. Y. Dong, "Online rare events detection," in Proc. PAKDD, Nanjing, China, 2007, pp. 1114–1121.
[9] P. Zhao, S. C. H. Hoi, R. Jin, and T. Yang, "Online AUC maximization," in Proc. 28th ICML, 2011, Bellevue, WA, USA, pp. 233–240.
[10] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," ACM CSUR, vol. 41, no. 3, Article 15, 2009.
[11] K. Crammer, O. Dekel, J. Keshet, S. Shalev-Shwartz, and Y. Singer, "Online passive-aggressive algorithms," J. Mach. Learn. Res., vol. 7, pp. 551–585, Mar. 2006.
[12] Y. Li and P. M. Long, "The relaxed online maximum margin algorithm," in Proc. NIPS, 1999, pp. 498–504.
[13] F. Rosenblatt, "The perceptron: A probabilistic model for information storage and organization in the brain," Psych. Rev., vol. 65, no. 6, pp. 386–408, 1958.
[14] M. Dredze, K. Crammer, and F. Pereira, "Confidence-weighted linear classification," in Proc. 25th ICML, Helsinki, Finland, 2008, pp. 264–271.
[15] P. Domingos, "Meta cost: A general method for making classifiers cost-sensitive," in Proc. 5th ACM SIGKDD Int. Conf. KDD, San Diego, CA, USA, 1999, pp. 155–164.
[16] C. Elkan, "The foundations of cost-sensitive learning," in Proc. 17th IJCAI, San Francisco, CA, USA, 2001, pp. 973–978.
[17] C. Gentile, "A new approximate maximal margin classification algorithm," J. Mach. Learn. Res., vol. 2, pp. 213–242, Dec. 2001.
[18] S. C. H. Hoi, R. Jin, P. Zhao, and T. Yang, "Online multiple kernel classification," Mach. Learn., vol. 90, no. 2, pp. 289–316, 2013.
[19] M. Sokolova, N. Japkowicz, and S. Szpakowicz, "Beyond accuracy, f-score and ROC: A family of discriminant measures for performance evaluation," in Proc. AAAI, Hobart, TAS, Australia, 2006, pp. 1015–1021.
[20] R. Akbani, S. Kwek, and N. Japkowicz, "Applying support vector machines to imbalanced datasets," in Proc. 15th ECML, Pisa, Italy, 2004, pp. 39–50.