Refined Clustering technique based on boosting and outlier detection

Ms. Reshma Y. Nagpure, Prof. P. P. Rokade

Abstract - Boosting is the repetitive process to perk up the accuracy in functions for prediction that supervised learning (SL) system learn using training data. In this prediction process, boosting considers multiple function rather that considering only single function from the same supervised learning system. Boosting process then predicts the label for new data instances using a weighted vote over all the functions. By considering and merging multiple functions together , boosting manage to get fine grained decision boundary on training data than using single function. Boosting for supervised learning having certain limitations like e.g. because of problematic data difficulty arises to analyze the data , over-fitting of training data , wrong label prediction by initial function etc. Previous worked reflected that boosting is resistant to over fitting problem. Also in case of wrong label prediction from function, boosting achieves higher accuracy when multiple functions are used to decide the labels for clusters. Previous work have some difficulties like A] Wrong data i.e. label noise in training data which causes wrong output instances and B] Another problem is that when feature of label instances are different and not relevant with respective rest of training data then its proper cluster cannot be defined properly. Hence there must be proposed system that work on these problems. Also clustering can be achieved on problematic dataset also. For this cluster based boosting (CBB) approach should be adopted to achieve this. Also along with CBB, the outlier detection should be achieved so that data will be easy to analyze and cluster can be formed smartly.

Keywords - Boosting, supervised learning, cluster based boosting, over fitting

1 Introduction

Boosting is the process used for machine research algorithms. Boosting is used to enhance the accuracy. It works with many functions consecutively focusing on incorrect occurrence. In SL system boosting process convert weak learners to strong one. But boosting still having difficulties on certain data sets with assured types of ambiguous training data when difficult functions overfit the data. In reality, overfitting is defined as some degree of errors in data, which attempts to make model to accommodate too closely fit to a limited set of data points. Specifically, the process of boosting learns many functions from SL system[2][4].In SL problem degree of the training data is overspecializes to produced perfect accuracy in new data.

In addition, boosting effectively used in wide range of application. Boosting also contains engineering machine[5] to predict concrete strengths[6].

In this paper, we proposed CBB to address the limitations of boosting. Cluster-based boosting (CBB) is referred to absorbs clusters into the boosting process. we show that how CBB determines the subsequent functions to enhance boosting. In cluster based boosting problematic training data contained by the cluster is addressed and divided. Then each cluster is separately evaluated by CBB to noticed that problematic training data, for that it needs more selective boosting, it helps to learn subsequent functions. Furthermore, to allow some minimum complexes in subsequent functions which helps to reduce overfitting is propagated into boosting, Finally, for more comprehesive boosting that can fullfill with problematic training data is required for CBB subsequent functions, starting with all the cluster member not only for those suspected incorrect by the initial function.

Here, in this paper we show novel cluster-based boosting (CBB) path for locating the control in boosting for administering(machine) learning system.

There are two specific limitations of current boosting those are resulting from boosting focusing on incorrect training data in our approach: (1) Filtering for subsequent functions when the training data contains troublesome areas and/or label noise and (2) Overfitting in subsequent functions that are forced to learn on all the incorrect instances.

Above limitations are addressed as follows: For each cluster CBB mitigates filtering for subsequent functions by using the appropriate amount of boosting.

CBB is figure out in three ways: First we analyze CBB to AdaBoost, the most famous boosting algorithm. Second, we analyze CBB to a previous algorithm, PruneBoost, that uses clusters, as a preprocessing step to improve boosting [10]. We also calculate the CBB clusters in more fact to consider addiction and natures of CBB. Third, we calculate CBB to consistence boosting algorithms, BrownBoost[15] and AdaBoostKL[11], which also use collective boosting. The changeable amounts of label noise and annoying areas also contained by these data sets.

In CBB boosts further cluster are divided according to the structure provided using a huge as well as minimum

- *Reshma Y. Nagpure is currently pursuing masters degree program in computer engineering in Pune University, India, PH-9021828352. E-mail: nagpure.reshma89@gmail.com*
- *P. P. Rokade is with Computer Science and Engineering Department, Pune University, India. E-mail: prakashrokade2005@gmail.com*

learning rate, or not boosting on each cluster based previous function accuracy on the member data. Generally, predictive accuracy on problematic training data in CBB is allowed by boosting. In recent absorption, error bound that are heavily related with the whole margin distribution[4], for that boosting minimizes the function complexity to ease the overfitting. 20 UCI benchmark data set with three kinds of examining learning systems are generated global experimental result that are provided by us[4].

These results determine the capability of CBB access compared to a popular boosting algorithm, in that two algorithms that are used for selective boosting after clustering.

2 LITERATURE SURVEY

2.1 Cluster-Based Boosting

This paper proposed solution over problem occur in clustering when wrong labels in training dataset is present and irrelevant features of instance members are found as far as rest of dataset is concern.

2.2 Evidence Contrary to the Statistical View of Boosting: A Rejoinder to Responses

In this paper firstly, in many researches the statistical community has concentrated on the former, for that an original AdaBoost algorithm was proposed. Secondly, with regard to minimization of misclassification error, the idea that boosting is reducing variance has been acknowledged in the discussions. In this authors define over fitting as a positive slope for a specified loss metric as a function of the iterations. Specifically, the loss metric they focus on is misclassification error.

2.3 On the doubt about margin explanation of boosting

In this paper, author primary focused on representation of the *kth margin bound* and further study on its relationship. They maintained the margin-based explanation against Breiman's doubts by supporting a new generalization error bound that is heavily related to the whole margin distribution. Authors also studied about the margin distribution bounds for generalization error of voting classifiers in finite VC-dimension space.

2.4 A review of data mining applications for quality improvement in manufacturing industry

This paper proposed several analyses on selected quality tasks are provided on DM applications in the manufacturing industry. Author reviewed quality tasks as product/process quality description, predicting quality, classification of quality, and parameter optimization. In data mining practices they focused on DM applications for each quality task and also for result of application is also examined.

2.5 Optimizing the prediction accuracy of concrete compressive strength based on a comparison of data-mining techniques

This paper focuses on data mining methods that are used to optimize accuracy of concrete (HPC).Modeling the dynamics of HPC is extremely challenging because that is highly complex. In this study the quantitative analyses were performed with the help of five different data mining methods: From that two are machine learning models, first one is artificial neural networks, and the other is support vector machines, third is statistical model (multiple regression), and remaining two are meta-classifier models i.e. multiple additive regression trees and bagging regression trees. As MART based modeling is effective for predicting the strength of varying HPC age, analytical results are based on MART modeling.

2.6 A data-mining approach to monitoring wind turbines

This paper researched about data-mining algorithms that are applied to build prediction models for wind turbine faults. Prediction process is divided into three-stage as followed: 1) Prediction of faulty kind.
2) Faults of the system of particular prediction.3) Unseen faults determination. An analysis of different data mining algorithms are reported that are based on data collected at a large farm. In this random forest algorithm models provided the best accuracy among all algorithms tested.

2.7 Avoiding boosting over fitting by removing confusing samples

In this paper proposed an algorithm for removing confusing samples and experimentally study behavior of AdaBoost trained on the resulting data sets. In result, that assured removing confusing samples helps in boosting process to reduce the generalization error and also it helps to avoid overfitting. Based on the work with the training sets process of removing confusing samples also provides an accurate error prediction.

2.8 Robust multiple manifolds structure learning

This paper introduces (RMMSL) scheme i.e. Robust Multiple Manifolds Structure Learning. This technique is used to estimate data structures robustly. In the local learning stage, RMMSL efficiently estimates local tangent space by weighted shallow-rank matrix factorization. The proposed a robust manifold clustering method is based on local structure learning results. By introducing a novel curved-level similarity function clustering method is designed to get the flattest manifolds clusters. They also demonstrate the effectiveness of the proposed approach, which yields higher clustering accuracy.

2.9 A survey on multi-view learning

In this paper, an author proposes multi-view learning approaches.
Authors review a number of representative multi-view learning algorithms in different areas that are classified them into three groups: 1) Co-training, 2) Multiple kernel learning, and 3) Subspace learning.

The base of these multiple –view learning is simultaneously accessing the multiple view. An exception of study is that learning a model from multiple views; it helps to do the study of how to form multiple views and manipulate these views. In this process consistency and complementary properties of different views are explored.

It has better generalization ability than single-view learning.

2.10 A Survey on Transfer Learning

In this paper, proposed survey focuses on categorizing and reviewing the current progress on transfer learning for classification, regression, and clustering problems.
This paper focuses on problems related to the data mining that are used on transfer learning for allocation, regression, and clustering problems that are related to data mining tasks.
Transfer learning is differentiating into three setting:
1. Inductive transfer learning,
2. Transductive transfer learning, and
3. Unsupervised transfer learning.

2.11 Resampling or reweighting: A comparison of boosting implementations

In this work, authors evaluate the boosting implementations analytically using imbalanced training data. With the help of 10 boosting algorithms, 4 learners and 15 datasets, they may find boosting by re-sampling over boosting re-weighting. Therefore authors can conclude that in general, boosting by re-sampling is preferred over boosting by weighting.

2.12 Bagging and AdaBoost algorithms for vector quantization

In this paper, authors proposed VQ methods. VQ is based on ensemble learning algorithms Bagging and AdaBoost. These methods contain one or more weak learner for training in parallel or sequentially. In bagging algorithm, the selected weak learners are trained in parallel from a given data set. Result of bagging gives the fair from the weak learners. And the result for AdaBoost is given as the weighted average among the weak learners. The presented simulation results show that the proposed methods can achieve a good performance in shorter learning times than conventional ones such as K-means and NG.

2.13 Clustering by learning constraints priorities

This paper proposes a method for creating a constrained clustering ensemble. This method learns the pair wise constraints with their priority. This paper results that the proposed method exceeds the original constrained K-means. This method integrates multiple clusters produced by using a simple constrained K-means algorithm that author modify to utilize the constraints priorities. The cluster ensemble is executed according to a boosting framework; it adaptively learns the priority of constraints.

2.14 Generic object recognition with boosting

In this paper authors presents whole framework that begin with the extraction of various local regions of either discontinuity. For each local region, variety of local descriptors can be applied to form a set of feature vectors. In this boosting is used to determine a subset of such uncertain hypotheses and to combine them into one final vector for each visual class. This paper result obtains allocation results up to 81 percent ROC-equal error rate on the most complex of their databases.

2.15 The WEKA Data Mining Software: An Update

This paper introduces WEKA workbench, reviews the history of the project. WEKA mainly concentrates on to provide large collection of machine learning algorithms and some data processing tools for researchers. It helps to new data sets to compare different machine learning methods.

2.16 Consequences of variability in classifier performance estimates

This paper compares stability and similarity of the algorithms. This is a methodological choice which may have significant impact .It result also includes statistical tests. In this particularly, authors were examining the performance metrics and data sets used, cross-validation employed, and the number of iterations of cross-validation run has a significant, and often predictable, effect.

2.17 Outlier Detection technique

In this paper some extreme members are detected and then they are matched with relevant clusters.

3 Proposed System

Based on the literature survey it is clear that in existing system there is problem of messed training dataset , wrong prediction due to irrelevant instance feature etc. Hence there must be proposed system that can work with problematic dataset having wrong labels. Also proposed system should work well in case there in no relevant features found in some candidate instances as far as rest of training dataset is concern. Proposed system should help to create refines clusters that has members having similar to each other and as different as possible from other cluster members. Also proposed system should work on outlier

detection so that non performing members or outlier member can fit into proper clusters.

4 CONCLUSION

Cluster based boosting (CBB) approach will be good for problematic dataset in which wrong labels and irrelevant features of instances are there. Cluster based boosting helps to refine the cluster boundaries so that relevant members can be together. Also contribution like outlier detection is necessary which helps to normalize the clusters and relevant members should be together.

ACKNOWLEDGMENT

I am glad to express my sentiments of gratitude to all who rendered their valuable guidance to me. I would like to express my appreciation and thankful to Prof. S. R. Durugkar, Head of Department, Computer Engineering, S.N.D. College of Engineering. and Research Center, Nashik. I am also thankful to my guide P. P. Rokade, Head of Department of IT Engineering, S.N.D. College of Engineering. and Research Center, Nashik. I thank the anonymous reviewers for their comments

REFERENCES:

[1]"Cluster-Based Boosting", L. Dee Miller and Leen-Kiat Soh, Member, IEEE.1–430.

[2] L. Reyzin and R. Schapire, "How boosting the margin can also boost classifier complexity," in Proc. Int. Conf. Mach. Learn., 2006, pp. 753–760.

[3] W. Gao and Z-H. Zhou, "On the doubt about margin explanation of boosting," Artif. Intell., vol. 203, pp. 1–18, Oct. 2013.

[4] G. K€oksal, I. Batmaz, and M. C. Testik, "A review of data mining applications for quality improvement in manufacturing industry," Expert Syst. Apps., vol. 38, pp. 13488–13467, 2011.

[5] J. Chou, C. Chiu, M. Farfoura, and I. Al-Taharwa, "Optimizing the prediction accuracy of concrete compressive strength based on a comparison of data-mining techniques," J. Comp. Civil Eng., vol. 25, pp. 242–253, 2011.

[6] A. Kusiak and A. Verma, "A data-mining approach to monitoring wind turbines," IEEE Trans. Sustainable Energy, vol. 3, no. 1, pp. 150–157, Jan. 2012.

[7] A. Vezhnevets and O. Barinova, "Avoiding boosting overfitting by removing confusing samples," in Proc. Eur. Conf. Mach. Learn., 2007, pp. 430–441.

[8] D. Gong, X. Zhao, and G. Medioni, "Robust multiple manifolds structure learning," in Proc. Int. Conf. Mach. Learn., 2012, pp. 321– 329.

[9] C. Xu, D. Tao, and C. Xu, "A survey on multi-view learning," CoRR, vol. 1304, pp. 1–59, 2013.

[10] S. Pan and Q. Yang, "A survey on transfer learning," IEEE Trans.Knowl. Data Eng., vol. 22, no. 10, pp. 1345–1359, Oct. 2010.

[11] C. Seiffert, T.Khoshgoftaar, J.Hulse, and A.Naplitano, "Resampling or reweighting: A comparison of boosting implementations," in Proc. IEEE Int. Conf. Tools Artif. Intell., 2008, pp. 445–451.

[12] N. Shigei, H. Miyajima, M. Maeda, and L. Ma, "Bagging and adaboost algorithms for vector," Neurocomputing, vol. 73, pp. 106–114, 2009.

[13] M. Okabe and S. Yamada, "Clustering by learning constraints priorities," in Proc. Int. Conf. Data Mining, 2012, pp. 1050–1055.

[14] A. Opelt, A. Pinz, M. Fussenegger, and P. Auer, "Generic object recognition with boosting," IEEE Trans. Pattern Anal. Mach. Intell., vol. 28, no. 3, pp. 416–431, Mar. 2006.

[15] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: An update," SIGKDD Explorations, vol. 11, pp. 10–18, 2009.

[16] T. Raeder, T. R. Hoens, and N. V. Chawla, "Consequences of variability in classifier performance estimates," in Proc. Int. Conf. Data Mining, 2010, pp. 421–430.

[17] Hans-Peter Kriegel, Peer Kröger, Arthur Zimek , "Outlier Detection Technique" The 2010 SIAM International Conference on Data Mining

Authors

Ms. Reshma Y. Nagpure received the B.E. degree in Information Technology from Sanjivani College of engineering, Kopargaon in 2013. She is currently pursuing her Masters degree in Computer Engineering from S.N.D. College of Engineering and Research Centre, Savitribai Phule Pune University Former UOP. This paper is published as a part of the research work done for the degree of Masters.

Prof. P. P. Rokade is HOD in Information Technology Engineering, S.N.D. College of Engineering and Research Centre, Savitribai Phule Pune University.