International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Impact Factor (2012): 3.358

Survey of Novel Method for Online Classification in Data Mining

Jadhav Bharat S[1], Gumaste S.V.[2]

[1]Pune University, Sharadchandra Pawar College of Engineering, Dumbarwadi, Otur, Pune, Maharashtra, India

[2]Pune University, Sharadchandra Pawar College of Engineering, Dumbarwadi, Otur, Pune, Maharashtra, India

Abstract: ***Nowadays in communities of Data Mining and Machine Learning, cost-sensitive classification and online learning have been widely examined. Even though these topics are getting more and more attention, very few studies are based on an important concern of Cost-Sensitive Online Classification. This problem can be explored widely and new technique can be implemented to deal with this issue. By directly enhancing cost-sensitive measures utilizing online gradient descent methods, a new technique can be implemented. Particularly, two novel cost-sensitive online classification algorithms can be presented, which are intended to specifically enhance two well-known cost-sensitive measures: (a) maximizing weighted sum of specificity and sensitivity, (b) minimizing weighted misclassification cost. The hypothetical limits of the cost-sensitive measures made by the algorithms should be examined. Also their experimental performance on number of different cost-sensitive online classification tasks should be examined. This technique can be efficiently utilized for solving number of online anomaly detection tasks. Also it is very efficient and effective technique to handle cost-sensitive online classification tasks in number of application domains.***

Keywords: Cost Sensitive Classification, Online Classification, Online Gradient Descent, Misclassification Cost.

1. Introduction

Today, a critical need in data mining and machine learning is to implement proficient and versatile algorithms for mining substantial fast developing data. A hopeful approach is to explore Online Learning, a family of proficient and adaptable machine learning techniques, which have been keenly, examined in literature [1] [2] [3]. Generally, the objective of online learning is to incrementally realize few prediction models to make accurate predictions on a stream of samples that arrive consecutively. Online learning is beneficial for its high proficiency and adaptability for extensive scale applications, and has been utilized to deal with online classification tasks in a number of different data mining applications. Different online learning systems have been proposed in literature. For example: the very popular Perceptron algorithm [3], Passive-Aggressive (PA) learning [1], and numerous other as of late proposed algorithms [4] [5] [6] [7]. Even though being studied widely, most of existing algorithms failed to handle the cost-sensitive classification tasks. The important issue in data mining which should be explored is misclassification costs [15] [16]. The current online learning methods are not successful enough due to the fact that most of existing online learning research worry about the performance of an online classification algorithm regarding prediction mistake rate, which is clearly taken a cost-insensitive and hence improper for a lot of real applications in data mining, particularly for cost-sensitive classification where datasets are frequently class-imbalanced and the misclassification costs of cases from distinctive classes can be extremely different [17] [18].

Researchers have presented many metrics to deal with problem of cost-sensitive classification. For Example: the weighted sum of sensitivity and specificity [18] and the weighted misclassification cost [16]. Form last decade, significant research has been done in order to develop batch classification algorithms for enhancing the cost-sensitive measures. But these algorithms suffer from inefficiency and poor scalability.

In communities of Data Mining and Machine Learning, cost-sensitive classification and online learning have been widely examined. Even though these topics are getting more and more attention, very few studies are based on an important concern of Cost-Sensitive Online Classification.

2. Literature Review

T. Yang et al. [4] explored new challenge, "Online Multiple Kernel Classification" (OMKC), which intends to attack an online learning task by learning in a kernel based prediction function from a pool of predefined kernels. To tackle this problem, they propose a novel schema by consolidating two sorts of online learning algorithms, i.e., the Perceptron algorithm that takes in a classifier for a given kernel, and the Hedge algorithm that consolidates numerous kernel classifiers by linear weighting. The solution to an OMKC task is an appropriate selection technique to pick a set of kernels from the group of predefined kernels for online classifier overhauls and classifier combination towards prediction. To address this key issue, they exhibit two sorts of selection technique:

a) Deterministic approach that picks the all kernels,
b) Stochastic approach that arbitrarily samples a subset of kernels as per their weights. Particularly, they proposed four variations of OMKC algorithms by implementing distinctive online updating and combination technique. Each of these four OMKC algorithms has distinctive benefits for diverse situations.

To inspect the observational performance of the presented OMKC algorithms, they carried out broad experiments a testbed with 15 different true datasets. The hopeful results uncover three real findings:

a. all of OMKC algorithms dependably perform better than a consistent Perceptron algorithm with an unbiased linear combination of various kernels, basically perform better than the Perceptron algorithm with the best kernel detected by validation, and frequently perform preferable or least equivalently over a state-of-the-art online MKL algorithm;
b. For the two diverse updating systems, the stochastic updating methodology has the capacity of fundamentally enhancing the effectiveness by keeping up at least comparable performance as contrasted and the deterministic methodology;
c. for the two distinctive combination methods, the deterministic combination method normally carry out better results, while the stochastic combination procedure has the capacity of producing an altogether more sparse classifier.

J. Wang et al. [5] proposed the Soft Confidence-Weighted (SCW) learning, another second-order online learning system with state-of-the-art experimental performance. Not at all like the current second-order algorithms, has SCW appreciated all the four properties.

a) Extensive margin training.
b) Adaptive margin.
c) Confidence weighting.
d) Ability of handling non-separable data.

Experimentally, they discovered the proposed SCW algorithms perform altogether better than the first CW algorithm, and beat the state-of-the-art AROW algorithm for most cases as far as both accuracy and proficiency.

S. C. H. Hoi et al. [8] had presented a Library for Online Learning Algorithms, which is called as LIBOL. LIBOL is a simple to utilize open source package for online learning development and study. The current interpretation of LIBOL consists of a substantial number of online learning algorithms for online-classification tasks. LIBOL is even now being enhanced by feedback from practical clients or actual users. They would like to make LIBOL a helpful machine learning instrument, as well as a perfect learning stage to carry out online learning examination. A definitive objective is to make simple learning with enormous data streams for handling the challenge of huge data analytics.

R. Jin et al. [9] examined another research issue of Online Feature Selection (OFS), which intends to select a predetermined number of features for prediction by an online learning design. They introduced a novel OFS algorithm to resolve the learning task, and offered hypothetical examination on the mistake bound of the proposed OFS algorithm. They broadly analyzed their observational performance on both standard UCI datasets and substantial scale datasets. They likewise thought about the proposed online feature selection system with a standard state-of-the-art characteristic selection algorithm for tackling real-world applications: picture characterization in computer vision and microarray gene interpretation examination in bioinformatics. Empowering results demonstrate the proposed algorithms are genuinely viable for feature selection tasks of online applications, and essentially more proficient and scalable than some state-of-the-art characteristic selection system.

To fill the space between cost-sensitive classification and online learning in machine learning and data mining, J. Wang, S. C. H. Hoi and P. Zhao [11] examined another methodology of Cost-Sensitive Online Classification, which means to specifically optimize cost-sensitive measures for online classification tasks. They proposed a group of compelling algorithms focused on online gradient descent, hypothetically investigated their cost-sensitive limits, and lastly analyzed their experimental execution widely. Their empowering results demonstrate that the proposed algorithms significantly do better than the conventional online learning algorithms for cost-sensitive online classification tasks. Through this study, they want to inspire research in both data mining and machine learning to further investigate top to bottom hypothesis of cost-sensitive online classification and the application of new cost-sensitive online learning procedures to handle a different types of developing difficulties in real-world data mining applications.

P. Zhao [12] researched another technique of Cost-Sensitive Online Classification that straightforwardly optimizes some cost-sensitive metrics. Particularly, they proposed two successful cost-sensitive DUOL algorithms focused on the current Double Updating Online Learning (DUOL) systems, hypothetically investigated their expense touchy limits, and lastly analyzed their observational performance widely, including their applications to online anomaly discovery tasks. Their empowering results demonstrate that their algorithms do better than the current algorithms for cost-sensitive online classification tasks.

In the study [13], they proposed a novel structure of Cost-Sensitive Online Active Learning (CSOAL) as a common, basic yet genuinely powerful approach to handling a real-world online malicious URL identification task. They displayed the CSOAL algorithms to advance cost-sensitive measures and hypothetically examine the limits of the proposed algorithms. They additionally widely analyzed their experimental performance on expansive scale real-world dataset. Their empowering results demonstrated that

a. The presented CSOAL system has the capacity extensively beat various supervised cost-sensitive online learning algorithms for malicious URL identification task;
b. The presented CSOAL system has the capacity achieve the equivalent or surprisingly better state-of-the-art predictive performance of a cost-sensitive online learner by questioning an altogether little measure of named information; and (c)the presented CSOAL algorithms are very proficient and adaptable for web-scale applications.

R. Jin, S. C. H. Hoi, and P. Zhao [6] exhibited a novel "Double Updating" methodology to online learning named as "DUOL". Double Updating Online Learning (DUOL) not just modifies the weight of one current support vector that the most sincerely conflicts with the new help vector, additionally updates the weight of the misclassified illustration. They demonstrate that the mistake bound for an online-classification task can be essentially diminished by the proposed DUOL algorithms. They have directed a broad set of investigations by algorithms with various algorithms for both twofold and multiclass online classifications. Hopeful experimental results demonstrated that the proposed double updating online learning algorithms reliably outflank the single-update online learning algorithms.

3. Proposed System

In this paper, issue of Cost-Sensitive Online Classification by endeavoring to create cost-sensitive algorithms for tackling a cost-sensitive classification task is considered and explained. In this paper, novel approach of Cost-Sensitive Online Classification to solve this problem is proposed. The main problem is the means by which to implement a successful cost-sensitive online algorithm which can enhance a predefined cost-sensitive measure for an online classification task, and further offer hypothetical assurance of the proposed algorithm.

This Problem can be resolved by

a) For handling the online optimization task of maximization of the weighted sum or minimization of the weighted misclassification cost, propose two cost-sensitive online learning algorithms utilizing online gradient descent system.
b) For cost-sensitive online classification tasks, hypothetically investigate the cost-sensitive measure limits of the proposed algorithms, and widely analyze their experimental performance for cost-sensitive online classification tasks. To deal with a data mining application that is online anomaly detection tasks, apply presented method.

4. Conclusion

Cost-Sensitive Classification and Online Learning have been studied widely in Data Mining and Machine Learning communities but problem of Cost-Sensitive Online Classification is still ignored. As an endeavor to fill the crevice between cost-sensitive classification and online learning, I explored another schema of Cost-Sensitive Online Classification to resolve focused around online gradient descent techniques. Then hypothetically examined their cost-sensitive bounds, further inspected their experimental performance, and lastly exhibited their applications to handle real-world online anomaly detection tasks. The empowering results of presented algorithm demonstrated that system accomplished the state-of-the-art performance for cost-sensitive online classification tasks.

References

[1] 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.
[2] Y. Li and P. M. Long, “The relaxed online maximum margin algorithm,” in Proc. NIPS, 1999, pp. 498–504.
[3] 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.
[4] 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.
[5] J. Wang, P. Zhao, and S. C. H. Hoi, “Exact soft confidence-weighted learning,” in Proc. 29th ICML, Edinburgh, U.K., 2012.
[6] P. Zhao, S. C. H. Hoi, and R. Jin, “Double updating online learning,” J. Mach. Learn. Res., vol. 12, pp. 1587–1615, May 2011.
[7] M. Dredze, K. Crammer, and F. Pereira, “Confidence-weighted linear classification,” in Proc. 25th ICML, Helsinki, Finland, 2008, pp. 264–271.
[8] S. C. H. Hoi, J. Wang, and P. Zhao, “LIBOL: A library for online learning algorithms,” J. Mach. Learn. Res., vol. 15, no. 1, pp. 495–499, 2014.
[9] S. C. H. Hoi, J.Wang, P. Zhao, and R. Jin, “Online feature selection for mining big data,” in Proc. 1st ACM Int. Workshop BigMine, Beijing, China, 2012, pp. 93–100.
[10] 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.
[11] J. Wang, P. Zhao, and S. C. H. Hoi, “Cost-sensitive online classification,” in Proc. 12th IEEE ICDM, Brussels, Belgium, 2012.
[12] P. Zhao and S. C. H. Hoi, “Cost-sensitive double updating online learning and its application to online anomaly detection,” in Proc. SDM, Austin, TX, USA, 2013.
[13] P. Zhao and S. C. H. Hoi, “Cost-sensitive online active learning with application to malicious URL detection,” in Proc. 19th ACM SIGKDD Int. Conf. KDD, Chicago, IL, USA, 2013.
[14] P. Domingos, “Metacost: A general method for making classifiers cost-sensitive,” in Proc. 5th ACM SIGKDD Int. Conf. KDD, San Diego, CA, USA, 1999, pp. 155–164.
[15] C. Elkan, “The foundations of cost-sensitive learning,” in Proc. 17th IJCAI, San Francisco, CA, USA, 2001, pp. 973–978.
[16] H. Masnadi-Shirazi and N. Vasconcelos, “Risk minimization, probability elicitation, and cost-sensitive svms,” in Proc. 27th ICML, Haifa, Israel, 2010, pp. 759–766.
[17] C. Drummond and R. C. Holte, “C4.5, class imbalance, and cost sensitivity: Why under-sampling beats over-sampling,” in Proc. ICML, 2003, pp. 1–8.
[18] 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.

Author Profile

Mr. Bharat S. Jadhav received the BE degree in Information Technolgy from Pravara Rural Engineering College in 2012 . During 2013-2014, he stayed in Late Hon.D.R.Kakade Polytechnic Pimaplwandi as lecturer in Computer Technology Department,. He now study Master Of Engineering in Sharadchandra Pawar College of Engineering, Dumbarwadi,Otur, University Of Pune , Inc.

Mr. S.V.Gumaste, currently working as Professor and Head, Department of Computer Engineering, SPCOE-Dumberwadi, Otur. Graduated from BLDE Association's College of Engineering, Bijapur, Karnataka University, Dharwar in 1992 and completed Post-graduation in CSE from SGBAU, Amravati in 2007 and submitted thesis on "Study and Analysis of Optimization of Band width parameters in Adhoc Networks" for award of Degree of Ph.D (CSE) in Engineerng & Faculty at SGBAU, Amravati. Has around 22 years of Teaching Experience.