Feature selection for multiclass classification and regression techniques

N. Lakshmi[1], N. Deepika[2]

Asst. Professor, Computer Science & Engineering, DMI College of Engineering, Tamilnadu, India[1,2]

Abstract: Online Multiclass Classification (OMC) performs the heterogeneous domain from complex data of completely diverse feature representation. OMC algorithm investigates the problem of heterogeneous domain and regression problems. Most existing studies of online learning divide the online feature selection into two parts. i) Learning with full input and ii) Learning with partial input. To address this limitation, we investigate the heterogeneous and regression problem in which an online learner is allowed to maintain a classifier with limited number of features. The key challenge of OMC algorithm is how to maintain the multiclass classification and regression using the active features. We attempt to tackle this challenge by studying on line feature selection and truncation techniques. We present OMC, Novel algorithm to solve the problems and give their performance analysis. We evaluate the performance of the proposed algorithms for on line learner using different domain and demonstrate their applications in real world problems including image classifications and analysis of bio informatics. Encouraging results validate the efficiency of our techniques.

Keywords: online multiclass classification, online Learning, Large scale data mining, Big data analytics.

1. INTRODUCTION

Transfer learning mainly aims to address the machine learning tasks of building models in a new target domain by taking advantage of information from another existing source domain through knowledge transfer. Unlike the existing transfer learning studies, in this paper, we propose a new framework of Online Multiclass Classification (OMC), which addresses the transfer learn-ing problem using an online learning framework. As the first attempt to this problem, we address some OMC challenges in two different settings. In the first setting, we study the homogeneous OMC where the target domain shares the same feature space as the old/source one.In the second setting, we address the challenge of heterogeneous OMC where the feature space of the target domain is different from that of the source domain. We propose algorithms to solve both problems, and theoretically analyze their mistake bounds. Finally, we empirically examine their perfor-mance on several challenging OMC tasks.

The major contributions of this paper include the following:

1. we propose novel algorithms to solve both of the above OFS tasks;
2. we analyze theoretical properties of the proposed algorithms;
3. we validate their empirical performance by conduct-ing an extensive set of experiments;
4. finally, we apply our technique to solve real-world problems in text classification, computer vision, and bioinformatics.

2. RELATED WORKS

Unlike typical machine learning methods that assume training examples are available before the learning task, online learning is more appropriate for some real-world problems where training data arrive sequentially. Due to their merits of attractive efficiency and scalability, various online learning methods have been proposed. One well-known approach is the Perceptron algo-rithm [8],[9],which updates the model by adding a new example with some constant weight into the current set of support vectors when the example is misclassified. Recently many online learning algorithms have been proposed based on the criterion of maximum margin [1],[10].One example is the Passive-Aggressive (PA) method [1], which updates the classification model when a new example is misclassified or its classification score is smaller than some predefined margin.More extensive surveys for online learning can be found in [11]..

Transfer learning (TL) has been actively studied. The goal of TL is to extract knowledge from one or more source tasks and then apply them to a target task. Various TL methods have been proposed. Accord-ing to different learning types, these methods can be roughly classified into three categories: *induc-tive*, *transductive*, and *unsupervised* approaches. In-ductive TL [2], aims to in-duce the model in the target domain with the aid of knowledge transferred from the source domains; trans-ductive TL [4] aims to extract the knowledge from source domain to improve the predic-tion tasks in the target domain without labeled data in the target domain; while unsupervised TL aims to resolve unsupervised learning tasks in target do-main [6]. Moreover, according to dif-ferent feature representation, TL can be classified as *homogeneous* vs. *heterogeneous* TL [7] where the feature spaces of source and target domains can be different. A comprehensive survey on transfer learning can be found in [3].Although both online learning and transfer learning have been actively studied, to the best of our knowledge, no existing work has formally addressed transfer learning by an online learning framework. Finally, we note that OTL is also different

from online multi-task learning [5], which aims to learn multiple tasks in parallel in an online learning framework.

3. ONLINE MULTICLASS CLASSIFICATION

3.1 OMC: Homogeneous Domain

We start by studying the homogeneous OTL, in which we assume the source domain and the target domain have the same feature space, i.e., X2 = X1 and Y2 = Y1. One key challenge of this task is to address the *concept drifting* issue that often occurs in this scenario. Specifically, the concept drift means that the target variable to be predicted changes over time in the learning process. This raises the challenge of transferring knowledge from source domain to target domain. The basic idea of our OTL solution is based on the ensemble learning approach. In particular, we first construct an entirely new prediction function f only from the data in the target domain in an *online* fashion, and then learn an ensemble prediction function that is the mixture of both the old and the new prediction functions, i.e., h and f, which thus can transfer the knowledge from the source domain. The remaining issue is then how to effectively combine the two prediction functions for handling the concept drift issue.To combine the two prediction functions h(x) and ft(x) at the t-trial of the online learning task, we introduce two weight parameters, w1,t and w2,t, for the two prediction functions respectively. At the t-th step, given an instance x2t , we predict its class label by the following ensemble function:

3.2 OMC:Heterogeneous Domains

In this section, we study the OTL problem across heterogeneous domains where the source and target domains have different feature spaces. Heterogeneous OTL is generally very challenging. To simplify the problem, we assume the feature space of the source domain is a subset of that of the target domain. Due to the difference of the two feature spaces, we cannot directly apply the algorithm in the previous section. Below we propose to introduce a multi-view approach for solving the challenge in this case.

3.3 Simple Truncation Approach

A straightforward approach to online feature selection is to modify the perceptron algorithm by applying truncation. Specifically, In the t-th trial, when being asked to make prediction, we will truncate the classifier wt by setting everything but the B largest (absolute value) elements in wt to be zero. This truncated classifier, denoted by wBt , is then used to classify the received instance xt. Similar to the perceptron algorithm, when the instance is misclassified, we will update the classifier by adding the vector ytxt, where ðxt; ytÞ is the misclassified training example. Algorithm 1 shows the steps of this approach.

Algorithm 1. Modified Perceptron by Truncation for OFS.
1: Input
. B: the number of selected features
2: Initialization
. w1 ¼ 0
3: for t ¼ 1; 2; . . . ; T do
4: Receive xt
5: Make prediction sgnðx>t wtÞ
6: Receive yt
7: if ytx>t wt _ 0 then

8: bwtþ1 ¼ wt þ ytxt
9: wtþ1 ¼ Truncateðbwt�þ1;BÞ
10: else
11: wtþ1 ¼ wt
12: end if
13: end for
Algorithm 2. w ¼ Truncateðbw;BÞ.
1: if kbwk0 > B then
2: w ¼ bwB where bwB is bw with everything but the B largest elements set to zero.
3: else
4: w ¼ bw
5: end if

4. CONCLUSIONS

In this paper, we studied the new problem of Online Multiclass Classification (OMC), which aims to transfer knowledge from a source domain to an online learning task on a target domain. We addressed two OMC tasks in different settings and presented two novel OMC algorithms. We offered theoretical analysis on the mistake bounds of the proposed OMC algorithms, and extensively examined their empirical performance. Encouraging results show the proposed algorithms are effective. Through this work, we hope to encourage the investigation of OMC to address other harder problems ,e.g. how to perform heterogeneous OMC from complex data of completely diverse feature representations.

REFERENCES

[1]. Crammer, Koby, Dekel, Ofer, Keshet, Joseph, Shalev-Shwartz, Shai, and Singer, Yoram. Online passiveaggressive algorithms. J. Mach. Learn. Res., 7:551–585, 2006.
[2]. Daum´aIII, H. and Marcu, D. Domain adaptation for statistical classifiers. Journal of Artificial Intelligence Research, 26:101–126, 2006.
[3]. Pan, Sinno Jialin and Yang, Qiang. A survey on transfer learning. IEEE Transactions on Knowledge and Data Engineering, 2009.
[4]. Arnold, Andrew, Nallapati, Ramesh, and Cohen,William W. A comparative study of methods for transductive transfer learning. In Proc. 7th IEEE Int'l Conf. on Data Mining Workshops, pp. 77–82, Washington, DC, USA, 2007.
[5]. Dekel, Ofer, Long, Philip M., and Singer, Yoram. Online learning of multiple tasks with a shared loss. J.Mach. Learn. Res., 8:2233–2264, 2007.
[6]. Dai, Wenyuan, Yang, Qiang, Xue, Gui-Rong, and Yu,Yong. Self-taught clustering. In Proc. 25th Int'l Conf. on Machine Learning (ICML2008), pp. 200–207, Helsinki, Finland, 2008.
[7]. Argyriou, Andreas, Maurer, Andreas, and Pontil, Massimiliano. An algorithm for transfer learning in a eterogeneous environment. In Euro. Conf. Mach. Learn. and Knowledge Discovery in Databases, pp.71–85, Antwerp, Belgium, 2008.
[8]. Rosenblatt, F. The perceptron: A probabilistic model for information storage and organization in the brain. Psychological Review, 65:386–407, 1958.
[9]. Freund, Yoav and Schapire, Robert E. Large margin classification using the perceptron algorithm. Mach. Learn., 37(3):277–296, 1999.
[10]. Li, Yi and Long, Philip M. The relaxed online maximum margin algorithm. In Advances in Neural Information Processing Systems (NIPS), pp. 498–504,1999.
[11]. Shalev-Shwartz, Shai. Online learning:theory, algorithms, and applications. In Ph.D thesis, 2007.
[12]. Jialei Wang, Peilin Zhao, Steven C.H. Hoi," online feature selection and its applications",Vol 26, No.3, March 2014.

BIOGRAPHIES

N.Lakshmi MCA., M.Tech Currently working as an Asssistant Professor in DMI College of Engineering

N. Deepika M.Tech,
Currently working as an Asssistant Professor in DMI College of Engineering