Neurocomputing 162 (2015) 163–170

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Update vs. upgrade: Modeling with indeterminate multi-class active learning

Xiao-Yu Zhang [a], Shupeng Wang [a,*], Xiaobin Zhu [b,**], Xiaochun Yun [a], Guangjun Wu [a], Yipeng Wang [a]

[a] Institute of Information Engineering, Chinese Academy of Sciences, 89 Minzhuang Road, Haidian District, Beijing 100093, China
[b] Beijing Technology and Business University, China 11 Fucheng Road, Haidian District, Beijing 100048, China

ARTICLE INFO

Article history:
Received 17 November 2014
Received in revised form
10 March 2015
Accepted 19 March 2015
Communicated by Dr. Ran He
Available online 2 April 2015

Keywords:
Active learning
Multi-class classification
Selective sampling
Model update
Model upgrade

ABSTRACT

This paper brings up a very important issue for active learning in practice. Traditional active learning mechanism is based on the assumption that the number of classes happens to be known in advance, and thus selective sampling is confined to the determinate model. However, as is the case for many applications, the model class is usually indeterminate and there is every chance that the hypothesis itself is inappropriate. To address this problem, we propose a novel indeterminate multi-class active learning algorithm, which comprehensively evaluates the instance based on both the value in refining the existing model and the potential in triggering model rectification. In this way, balance is effectively achieved between model update and model upgrade. Advantage of the proposed algorithm is demonstrated by experiments of classification tasks on both synthetic and real-world dataset.

1. Introduction

Active learning [1,2] is a kind of well-motivated algorithm for many machine learning problems where unlabeled data are abundant or easily obtained but labeled data are scarce or expensive to obtain. The key idea behind active learning is that the algorithm can actively select the most informative instances to be labeled by an oracle. By means of human–computer interaction, active learning aims to achieve high learning performance using as few labeled instances as possible, thereby minimizing the cost of obtaining labeled data.

Active learning has been widely applied to various classification applications [3–7]. According to the number of classes that the instances are classified into, classification can be divided into two problems, i.e. binary classification and multi-class classification. The former only involves two classes, whereas the latter is the task of assigning an instance to one of more than two classes. Binary classification is an extensively used model, which finds its application in many typical applications. It is also a basic classification model, based on which many algorithms have been developed. Multi-class classification is typically thought of as a generalization of binary classification, whose implementation requires either the combination of multiple binary classifiers or explicit model extension. Selective sampling plays a critical part in active learning. Various active learning scenarios have been formulated with both binary and multi-class versions. For example, a most straightforward method is to select the instances about which the model is least certain how to label [8–11]. Given multiple models, the query-by-committee strategy can be used in a voting manner and selects the instance with the greatest disagreement [12–15]. Another more theoretically-motivated approach is based on the measurement on how much the model's generalization error is likely to be reduced, and selects the instance with minimal expected future error [16–19].

Most active learning algorithms proposed so far are based on the assumption that the number of classes is determinate in advance, which is named Determinate Multi-Class Active Learning (DMC-AL). For DMC-AL, the right hypothesis is made at the beginning, and the remaining task is to refine the model parameter with informative instances. Unfortunately, for many applications, the number of classes cannot be revealed based on prior knowledge, or even varies with time. In this case, there is every chance that the existing model is inappropriate. As a result, not only the parameter under the existing model should be tuned for optimization, but also the model itself needs to be rectified with

* Corresponding author. Tel.: +86 10 82546670
** Corresponding author.
E-mail addresses: zhangxiaoyu@iie.ac.cn (X.-Y. Zhang), wangshupeng@iie.ac.cn (S. Wang), brucezhucas@gmail.com (X. Zhu), yunxiaochun@iie.ac.cn (X. Yun), wuguangjun@iie.ac.cn (G. Wu), wangyipeng@iie.ac.cn (Y. Wang).

http://dx.doi.org/10.1016/j.neucom.2015.03.056
0925-2312/

the incorporation of new model class. Active learning in this context is a much more challenging task.

For clarity, we use "model update" to represent the refinement of an existing model, and "model upgrade" to denote the transmutation into a new type of model with more variety. The traditional DMC-AL only focuses on model update, whereas neglects the possibility of model upgrade. In this paper, we propose a novel Indeterminate Multi-Class Active Learning (IMC-AL) algorithm for classification with indeterminate number of classes. With a comprehensive evaluation of the instance's informativeness, IMC-AL effectively achieves balance between model update and upgrade.

2. Determinate multi-class active learning

In the text that follows, we let x denote the input feature of an instance, and $y \in C^N = \{1, \ldots, N\}$ denote the class label, where N is the number of classes. U and L stand for the unlabeled and labeled dataset, respectively. Probabilistic model is used for classification based on the posterior distribution $P(y|x; \theta_L^N)$ of label y conditioned on the input x, where θ_L^N is the N-class classification model parameter optimized for the corresponding labeled dataset L.

For determinate multi-class classification, the number of classes N is preset as a constant based on empirical analysis or prior knowledge. Classification model can then be constructed for the multiple classes, which fall into two scenarios according to the modeling mechanism adopted.

A straightforward approach is reducing the problem of multi-class classification to multiple binary classification problems [20]. In the training stage, binary classifiers are trained. Each classifier corresponds to a class or a pair of two classes, based on either one-vs.-rest or one-vs.-one reduction strategy. In the prediction stage, the multiple binary classifiers are combined to arrive at a classification decision in a voting or fusion manner. For example, for each class $c \in C^N$, logistic regression [21] can be implemented in the binary classification setting with label $y_c \in \{0, 1\}$ for an instance x indicating whether it falls into class c or not, and the posterior distribution can be represented by the hypothesis function as follows:

$$P(y_c = 1 \middle| x; \theta_{L_c}^2) = h_{\theta_{L_c}^2}(x) = \frac{1}{1 + \exp\left(-(\theta_{L_c}^2)^T x\right)}. \tag{1}$$

The predicted label can be assigned with the class achieves the highest posterior probability.

$$y^* = \arg\max_{c \in C^N} P(y_c = 1 \middle| x; \theta_{L_c}^2). \tag{2}$$

For all its simplicity, this binary fusion modeling is a heuristic that suffers from several problems. Firstly, the combination of multiple binary classifiers is no easy task. Outputs in the form of discrete class labels can lead to ambiguities where some regions of the input space may receive the same number of votes, meanwhile real-valued confidence scores may differ in the scale between the binary classifiers. Secondly, even if the class distribution is balanced in the training dataset, the binary classification learners following the one-vs.-rest strategy see unbalanced distributions because typically the set of negatives they see is much larger than the set of positives. Furthermore, since multiple classification models need to be trained and combined, the efficiency of binary fusion modeling is jeopardized, especially for the one-vs.-one strategy where $N(N-1)/2$ binary classifiers are required.

The other approach is multi-class modeling, which considers all the classes jointly and constructs the model as a whole [22]. For example, as an expansion of logistic regression, softmax regression [23] generalizes logistic regression to the multi-class classification setting. Given an input x, softmax regression estimates the probability of the class label y taking on each of the N different possible values with an N-dimensional vector generated from the hypothesis as follows:

$$\begin{bmatrix} P(y=1 \middle| x; \theta_L^2) \\ P(y=2 \middle| x; \theta_L^2) \\ \cdots \\ P(y=N \middle| x; \theta_L^2) \end{bmatrix} = h_{\theta_L^N}(x) = \frac{1}{\sum_{i=1}^N \exp\left((\theta_L^N)_i^T x\right)} \begin{bmatrix} \exp\left((\theta_L^N)_1^T x\right) \\ \exp\left((\theta_L^N)_2^T x\right) \\ \cdots \\ \exp\left((\theta_L^N)_N^T x\right) \end{bmatrix}. \tag{3}$$

Since all the classes are modeled simultaneously in a unified process, the outputs are directly comparable to arrive at the prediction and the class distribution is relatively more balanced. As a result, the multi-class modeling is more applicable to multi-class classification problems.

After model construction, active learning is implemented for model update by feeding the existing model with the most informative instances via human–computer interaction.

Following the expected error reduction criterion for selective sampling, the most informative instance is the one that maximizes the expected error reduction, or equivalently minimizes the expected entropy over unlabeled dataset. Using $\tilde{y}$ to denote each possible label that an instance x may take on, the most informative instance under the DMC-AL scenario is selected as follows:

$$x^*_{\text{DMC-AL}} = \arg\min_{x \in U} \sum_{\tilde{y} \in C^N} P(\tilde{y} \middle| x; \theta_L^N) F(x, \tilde{y}; \theta_L^N), \tag{4}$$

where

$$\begin{aligned} F(x, \tilde{y}; \theta_L^N) &= \sum_{x_u \in U - x} H(y_u \middle| x_u; \theta^N_{L+(x,\tilde{y})}) \\ &= \sum_{x_u \in U - x} \left(- \sum_{\tilde{y}_u \in C^N} P(\tilde{y}_u \middle| x_u; \theta^N_{L+(x,\tilde{y})}) \cdot \log P(\tilde{y}_u \middle| x_u; \theta^N_{L+(x,\tilde{y})}) \right), \end{aligned} \tag{5}$$

represents the sum of entropy along all the rest unlabeled instances $x_u \in U - x$, given the existing model parameter θ_L^N and the newly labeled instance $(x, \tilde{y})$. $H(y_u | x_u; \theta^N_{L+(x,\tilde{y})})$ is the conditional entropy of the label y_u with respect to the instance x_u, and $L + (x, \tilde{y})$ stands for the labeled dataset with a new instance $(x, \tilde{y})$ incorporated.

After selective sampling, the most informative instance $x^*_{\text{DMC-AL}}$ is labeled and incorporated into the training dataset to further update the existing model.

3. Indeterminate multi-class active learning

In this section, we give a detailed description of the proposed indeterminate multi-class active learning, along with a comparison with determinate multi-class active learning.

3.1. Preliminary

For indeterminate multi-class classification, the number of classes N is evaluated as that of unique labels in the current labeled dataset, and is constantly updated with the growth of the training dataset. For example, as illustrated in Fig. 1(1), the initial labeled dataset comprises only two points, i.e., A and B, based on which a binary classification model is constructed. When point C is labeled in Fig. 1(2), the classification boundary is refined. However, since no new label is incorporated, it is still a binary classifier. Only after the label of D is revealed, a new class label emerges and thus