Expert Systems With Applications 113 (2018) 531–543

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Feature ranking for enhancing boosting-based multi-label text categorization

Bassam Al-Salemi*, Masri Ayob, Shahrul Azman Mohd Noah

Center for Artificial Intelligence Technology, Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, Malaysia

ARTICLE INFO

Article history:
Received 24 January 2018
Revised 9 July 2018
Accepted 10 July 2018

Keywords:
RFBoost
Boosting
Multi-label learning
Text categorization
Feature ranking

ABSTRACT

Boosting algorithms have been proved effective for multi-label learning. As ensemble learning algorithms, boosting algorithms build classifiers by composing a set of weak hypotheses. The high computational cost of boosting algorithms in learning from large volumes of data such as text categorization datasets is a real challenge. Most boosting algorithms, such as AdaBoost.MH, iteratively examine all training features to generate the weak hypotheses, which increases the learning time. RFBoost was introduced to manage this problem based on a rank-and-filter strategy in which it first ranks the training features and then, in each learning iteration, filters and uses only a subset of the highest-ranked features to construct the weak hypotheses. This step ensures accelerated learning time for RFBoost compared to AdaBoost.MH, as the weak hypotheses produced in each iteration are reduced to a very small number. As feature ranking is the core idea of RFBoost, this paper presents and investigates seven feature ranking methods (information gain, chi-square, GSS-coefficient, mutual information, odds ratio, F1 score, and accuracy) in order to improve RFBoost's performance. Moreover, an accelerated version of RFBoost, called RFBoost1, is also introduced. Rather than filtering a subset of the highest-ranked features, FBoost1 selects only one feature, based on its weight, to build a new weak hypothesis. Experimental results on four benchmark datasets for multi-label text categorization) Reuters-21578, 20-Newsgroups, OHSUMED, and TMC2007(demonstrate that among the methods evaluated for feature ranking, mutual information yields the best performance for RFBoost. In addition, the results prove that RFBoost statistically outperforms both RFBoost1 and AdaBoost.MH on all datasets. Finally, RFBoost1 proved more efficient than AdaBoost.MH, making it a better alternative for addressing classification problems in real-life applications and expert systems.

1. Introduction

According to the International Data Corporation (Gantz & Reinsel, 2012), digital data on the Internet will grow to 40,000 exabytes by 2020, from 130 exabytes in 2005. These huge amounts of data are usually distributed over the World Wide Web in unstructured forms. Managing and organizing these data requires efficient and effective automatic text categorization systems. For this reason, text categorization is still an important research area that receives much attention in the research community and industry.

Text categorization involves automatically assigning texts to the appropriate categories (labels) from a set of predefined categories (Elghazel, Aussem, Gharroudi, & Saadaoui, 2016; Sebastiani, 2002). Many classification algorithms have been investigated for text categorization, such as naïve Bayes, k nearest neighbours (kNN), support vector machines (SVMs), and decision trees (Farid, Zhang, Rahman, Hossain, & Strachan, 2014; Jiang, Li, Wang, & Zhang, 2016; Onan, Korukoğlu, & Bulut, 2016; Pavlinek & Podgorelec, 2017; Trstenjak, Mikac, & Donko, 2014; Zhang, Liu, Zhang, & Almpanidis, 2017). However, these algorithms are restricted to single-label classification problems, in which each instance (each text, in our case) is assigned to only one class label. Yet by their very nature, texts may belong to more than one class (multi-label classification problem). For example, a news article about "education" may also relate to "economy" and/or "politics". Several multi-label classification algorithms have been proposed which extend the single-label classification algorithms to solve the multi-label problem, such as binary relevance (Boutell, Luo, Shen, & Brown, 2004), classifier chains (Read, Pfahringer, Holmes, & Frank, 2011), label powerset (Tsoumakas & Vlahavas, 2007), ranking by pairwise comparison (Hüllermeier, Fürnkranz, Cheng, & Brinker, 2008), calibrated ranking by pairwise comparison (Fürnkranz, Hüllermeier, Mencía, & Brinker, 2008), hierarchical embedding (Kumar, Pujari, Padmanabhan, Sahu, & Kagita, 2018), clustered intrinsic label correlations (Kumar et al., 2018) and label correlation exploitation algorithms (Yu, Pedrycz, & Miao, 2014).

* Corresponding author.
E-mail addresses: bassalemi@ukm.edu.my (B. Al-Salemi), masri@ukm.edu.my (M. Ayob), shahrul@ukm.edu.my (S.A.M. Noah).

https://doi.org/10.1016/j.eswa.2018.07.024
0957-4174/© 2018 Elsevier Ltd. All rights reserved.

AdaBoost.MH (Freund & Schapire, 1997), the multi-label version of AdaBoost (Schapire, Freund, Bartlett, & Lee, 1998), is accurate and considered to be one of the state-of-the-art multi-label classification algorithms. As a boosting algorithm, AdaBoost.MH iteratively builds a set of weak hypotheses and then combines them as a final classifier which is capable of estimating the multiple labels for a given instance. AdaBoost.MH uses binary features to generate the weak hypotheses of decision stumps. To build a weak hypothesis during a specific boosting round, AdaBoost.MH generates a set of weak hypotheses, equal in number to the training features. The weak hypothesis that minimizes the Hamming loss training error is then selected, and all other hypotheses are eliminated.

AdaBoost.MH's iterative examination of all the training features in its weak learning is time-consuming, particularly when the dataset is large (Esuli, Fagni, & Sebastiani, 2006). To address this limitation, Al-Salemi, Noah, and Ab Aziz (2016) introduced an improved version of AdaBoost.MH, named "RFBoost". RFBoost learns by first ranking the training features and then, during each boosting round, filtering and using a small subset of the top-ranked features to produce a new weak hypothesis. Experimental results show that RFBoost is a fast and accurate algorithm for multi-label text categorization. RFBoost's enhanced performance relative to AdaBoost.MH is due to its ranking of the training features: while AdaBoost.MH uses binary features to build its weak hypotheses, RFBoost uses weighted features. However, Al-Salemi, Ab Aziz, and Noah (2016) only investigated two feature ranking methods for RFBoost. One of these uses the conditional probability of the words across the labels obtained by labelled latent Dirichlet allocation (LLDA; Mcauliffe & Blei, 2007) as the features' weights. The other ranking method uses boosting weights obtained by executing *one* boosting round on the training set. Even though LLDA is an effective method for feature ranking, as a topic model it requires resampling the topics estimation, which may result in increased computation time for large volumes of data.

The aim of the present paper is twofold: to investigate several existing feature weighting methods, namely, information gain, chi-square, GSS-coefficient, mutual information, odds ratio, F1 score, and accuracy (Forman, 2003; Katrutsa & Strijov, 2017; Liu, Lin, Lin, Wu, & Zhang, 2017; Lu et al., 2017; Pascoal, Oliveira, Pacheco, & Valadas, 2017; Qian & Shu, 2015; Song, Jiang, & Liu, 2017), in order to improve RFBoost, and to propose an accelerated variant of RFBoost, named "RFBoost1". Feature weighting allows ranking features based on their weights. The proposed RFBoost1 selects only a single ranked feature, based on its weight, to pass to the base learner for generating a new hypothesis, which eliminates the need to examine all of the training features, as in AdaBoost.MH, or even a subset of the ranked features, as in RFBoost. An empirical analysis was also conducted to validate that RFBoost1 does not penalize the boosting theory; this is described in Section 4.3.

2. Preliminaries and problem statement

Given a training set of labelled documents $\boldsymbol{S} = \{(\mathrm{x}_1,\ Y_1),\ \ldots.,\ (\mathrm{x}_n,\ Y_n)\}$, where each document $\mathrm{x}_i \in \chi$ is assigned to a multiple category (label) Y_i: $Y_\mathrm{i} \subseteq \boldsymbol{Y}$, $\boldsymbol{Y} = (\boldsymbol{y_1},\ \ldots, \boldsymbol{y_m})$, let $\boldsymbol{T} = \{\boldsymbol{t_1}, \ldots,\ \boldsymbol{t_v}\}$ be the set of training terms extracted from $\boldsymbol{S}$. AdaBoost.MH infers a strong classifier (the final classifier) $H : \chi \ \times \boldsymbol{Y} \rightarrow \mathbb{R}$ from $\boldsymbol{S}$ as the combination of a set of weak hypotheses $(h^{(1)}(\mathrm{x}, y),\ \ldots,\ h^{(R)}(\mathrm{x}, y))$ with a small Hamming loss in the form $H(\mathrm{x}, y\ ) = \sum_{r=1}^{R}\ \alpha^{(r)} h^{(r)}(\mathrm{x}, y)$, where R is the number of rounds. A given document x is then assigned to a category y if and only if $H(\mathrm{x}, y)$ is positive.

To produce a new weak hypothesis $h^r(\mathrm{x})$ during a boosting round r, AdaBoost.MH examines all the training terms in $\mathcal{T}$ and determines their absence/presence in each document under each label to build a set of weak hypotheses $(h_1^{(r)}\ (\mathrm{x}), \ldots,\ h_v^{(r)}(\mathrm{x}))$, one for each term in $\mathcal{T}$. Then a single hypothesis at a certain term (called the "*pivot term*") that minimizes the Hamming loss is returned, and all other $(v-1)$ hypotheses are discarded. The examination of all training terms in each round increases the training time, especially when the data size is large.

The multi-label boosting algorithm RFBoost (Al-Salemi, Abaziz et al., 2016) controls the computational learning cost by first reducing the number of terms to be examined by means of feature ranking. Then, for each boosting round r, only a small subset of the ranked features of a fixed size k, which is a very small number compared to v (the number of training features), are filtered and used to build the weak hypothesis $h^{(r)}$. For the next boosting round $(r+1)$, the pivot term selected in round r is removed and replaced with the next ranked feature in the ranked feature list, and so on. An experimental analysis proved that RFBoost is faster and more accurate than AdaBoost.MH and all of the other algorithms that were examined in the evaluation.

Because the main factor accounting for RFBoost's good performance is its feature ranking, in the present paper we investigate several state-of-the-art feature weighting methods for ranking the features, in an attempt to improve RFBoost's performance. In addition, we propose a variant of RFBoost called "RFBoost1". Rather than filtering a subset of ranked features, as RFBoost does, RFBoost1 selects a single feature to pass to the base learner as a pivot term. This reduces the computational time for building one weak hypothesis from $O(nmv)$ in AdaBoost.MH, where n is the number of training documents, m is the number of labels, and v is the training vocabulary, to $O(nm1) = O(nm)$ in RFBoost1.

3. Related work

A simple approach to solving the multi-label classification problem involves transforming the multi-label task into a set of single-label subtasks. A single-label classifier is then used to solve each subtask, and the outputs are combined to solve the original multi-label task. To this end, methods such as binary relevance (Boutell et al., 2004), classifier chains (Read et al., 2011), label powerset (Tsoumakas & Vlahavas, 2007), ranking by pairwise comparison (Hüllermeier et al., 2008), and calibrated ranking by pairwise comparison (Fürnkranz et al., 2008) have been introduced and used to solve many multi-label classification problems. Despite their simplicity, transformation-based methods still depend on the single-label classifiers, and the huge number of single-label classifiers makes it difficult to decide which transformation methods count as state-of-the-art for multi-label classification. Furthermore, the transformation methods have been criticized for being time-consuming and exhaustive in terms of memory resources (Zhang & Zhou, 2014).

An alternative approach to solving the multi-label classification problem is to adapt a single-label algorithm to directly solve the multi-label problem. Several multi-label classifiers have been adapted from single-label classifiers. For example, multi-label *k*NN (ML*k*NN; Zhang & Zhou, 2007) was adapted from the traditional *k*NN algorithm for multi-label classification and uses the maximum posterior principle to assign a given test instance to a label based on the prior and posterior probabilities for labels' frequencies within the k nearest neighbours. Another multi-label classification algorithm adapted from the *k*NN algorithm, BR*k*NN, uses the binary relevance (BR) transformation with the *k*NN algorithm. However, BR*k*NN is more efficient because it reduces the number of the BR pairs for each label. In multi-label instance-based learning by logistic regression (IBLR-ML; Cheng & Hüllermeier, 2009), the *k*NN algorithm is combined with logistic regression and allows the interdependencies between class labels to be captured correctly, so that the multi-label classification is handled appro-