Knowledge-Based Systems 136 (2017) 15–26

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

A new Centroid-Based Classification model for text categorization

Chuan Liu[a], Wenyong Wang[a,b,*], Guanghui Tu[a], Yu Xiang[a], Siyang Wang[c], Fengmao Lv[a]

[a] School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China
[b] Shanghai Hefu Artificial Intelligence Technology (Group) Company Limited Hefu Institution of UESTC, Chengdu 611731, China
[c] Mathematics Department, University of California San Diego, La Jolla 92093, USA

ARTICLE INFO

Article history:
Received 13 January 2017
Revised 17 August 2017
Accepted 26 August 2017
Available online 30 August 2017

Keywords:
Text categorization
Centroid-Based Classifier
Machine learning
Gravitation Model

ABSTRACT

The automatic text categorization technique has gained significant attention among researchers because of the increasing availability of online text information. Therefore, many different learning approaches have been designed in the text categorization field. Among them, the widely used method is the Centroid-Based Classifier (CBC) due to its theoretical simplicity and computational efficiency. However, the classification accuracy of CBC greatly depends on the data distribution. Thus it leads to a misfit model and also has poor classification performance when the data distribution is highly skewed. In this paper, a new classification model named as Gravitation Model (GM) is proposed to solve the class-imbalanced classification problem. In the training phase, each class is weighted by a mass factor, which can be learned from the training data, to indicate data distribution of the corresponding class. In the testing phase, a new document will be assigned to a particular class with the max gravitational force. The performance comparisons with CBC and its variants based on the results of experiments conducted on twelve real datasets show that the proposed gravitation model consistently outperforms CBC together with the Class-Feature-Centroid Classifier (CFC). Also, it obtains the classification accuracy competitive to the DragPushing (DP) method while it maintains a more stable performance. Thus, the proposed gravitation model is proved to be less over-fitting and has higher learning ability than CBC model.

1. Introduction

Automatic Text Categorization (TC) (also known as text classification) is a task of assigning text documents to pre-defined categories. In recent years, TC has received more attention due to a large number of text documents available on the Internet, and it has been widely used in many applications such as text filtering [1] and web page recommendation [2].

Centroid-Based Classifier (CBC) [1,3–7] is one of the most popular TC methods. The basic idea of CBC is that an unlabeled sample should be assigned to a particular class if the similarity of this sample to the centroid of the class is the largest. Compared with other TC methods, CBC is efficient since its computational complexity is linear in the training phase, this merit is important for massive text classification task. Although it has been shown that CBC consistently outperforms other methods such as k-Nearest-Neighbors, Naive Bayes, and Decision Tree on a wide range of datasets [8], CBC often suffers from the model misfit [9,10] when the data is class-imbalanced distribution [11]. This class-imbalanced problem, that implies a high ratio of the number of instances in the majority class to the number of examples in the minority class, is common in many real problems in which the minority class is usually the one that has the highest interest from a learning task. The misfit with imbalanced datasets is that CBC is often biased towards the minority class and, therefore, there is a higher misclassification rate for the minority class and a lower recall rate for the majority class.

To solve the model misfit of CBC, numerous approaches have been proposed, such as Class-Feature-Centroid (CFC) [3], Generalized Cluster Centroid Based Classifier (GCCC) [12], DragPushing (DP) [6] and Large Margin DragPushing (LMDP) [7]. However, the existing improvements of CBC focus on the methods that aim to obtain good centroids in construction or training phase. Therefore, they can not solve the inherent disadvantages of CBC model.

In this paper, the authors believe that different classes should share different similarities to the unlabeled sample in the class-imbalanced dataset. Therefore, the proposed model motivated by Newton's law of universal gravitation concentrates on producing a

* Corresponding author at: School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China.
E-mail addresses: mrliuchuan@foxmail.com, wangwy@uestc.edu.cn (W. Wang).

http://dx.doi.org/10.1016/j.knosys.2017.08.020

Voronoi partition [13] that directly alleviates the class-imbalanced problem. Hence, the proposed method is to adjust the classification hyperplane instead of the location of centroids to improve the classification performance. In the proposed model, each class is weighted by a mass factor to indicate the data distribution of the corresponding class, and the value of mass factor can be learned from the training set. Then, a new document will be assigned to a particular class with the max gravitational force. The proposed method is empirically evaluated by comparing it with three frequently-used centroid-based methods (i.e., CBC, CFC, and DP) based on twelve famous datasets in the field of text categorization. The experimental results demonstrate that the proposed method works well on the real-world datasets, and outstandingly outperforms the centroid-based approaches (e.g., CBC and CFC). Furthermore, it maintains a more stable performance than the state-of-the-art method.

The remainder of this paper is organized as follows: Section 2 summarizes the related work. Section 3 describes the original CBC. Three representative variants of CBC are presented in Section 4. The proposed gravitation model is introduced in Section 5. The performance of gravitation model is evaluated in Section 6. The paper closes with the conclusions and future works in Section 7.

2. Related work

A growing number of statistical classification methods and machine learning techniques which include Naive Bayes (NB) [14], K-Nearest Neighbor (KNN) [15,16], Neural Network (NN) [17–21], Decision Tree (DT) [22], Support Vector Machines (SVM) [23,24], CBC [1,3–7,25,26], graph-based approach [27] and classifier ensemble approaches [28], have been applied to text categorization in recent years. Naive Bayes is an effective probability model for text mining tasks. However, the performance is poor in case there are no sufficient training documents for each category [14]. KNN is a type of instance-based learning algorithm, but it suffers from several drawbacks [29] such as high storage requirement, low efficiency in classification response, and low noise tolerance. Thus, it is inefficient in online classification [15]. In a neural network based TC, many studies exploited deep learning methods in recent years. Word embeddings are used to represent dense and low-dimensional real-valued vectors. Additionally, the embeddings are capable of solving sparsity problem. For instance, Mikolov et al. represent word2vec in [19–21] and also doc2vec/paragraph vector in [30]. These models have the potential to overcome the weaknesses (i.e., ignoring the ordering/semantics of the words) of bag-of-words models. Decision tree performs well when there is a small number of features. However, it becomes difficult to create a text classifier for a large number of features [22]. SVM has been proved to be a state-of-the-art classifier in TC filed. However, the training of SVM is a very slow process and has become a bottleneck, since the Quadratic Programming (QP) problem that implies high training time complexity $O(n^3)$ and space complexity $O(n^2)$ needs to be solved [31]. Thus, the weakness with SVM is that large-scale training instances will lead to slow learning speed and large buffer memory requirement. To reduce the high computational complexity of SVM, a feasible approach is to reconstruct training set for SVM in the context of the reduction methods [31,32]. The graph-based approach for text classification have proven to achieve good performances as demonstrated in [27]. Also, many classifier ensemble approaches such as bagging [33], boosting [34], random subspace [35], and random forest [36] have been successfully applied to different areas and have achieved excellent performances. However, these approaches are accompanied by poor efficiency.

Several investigations [37,38] indicate that standard learning algorithms suffer from the significant loss of performance in the imbalanced dataset scenario in classification. Therefore, many solutions which can be categorized into three major groups [39] i.e., data sampling [40,41], cost-sensitive learning [42,43] and ensemble techniques [44–46] have been proposed to deal with this problem. Data sampling is to produce a more balanced class distribution that allow the learning algorithms to perform in a similar manner to standard classification. Thus, the main advantage of resampling techniques is that they are independent of the underlying learning algorithms, and it has proved empirically to be an useful solution that applying a preprocessing step to balance the class distribution. Resampling techniques can be mainly categorized into two families, i.e., undersampling methods and oversampling methods, and the main difference between them is dependent on the object that they deal with. Undersampling methods are based on data cleaning techniques to create a subset of the original dataset by eliminating the instances of majority class. A wide variety of undersampling methods such as Wilsonâs edited nearest neighbor (ENN) [47] rule, the one-sided selection (OSS) [48], the condensed nearest neighbor rule (CNN) [49] and the neighborhood cleaning rule [50] have been proposed. Conversely, the main idea of oversampling methods is to interpolate several minority class instances into the original dataset by replicating some instances or creating new instances from existing ones. Some representative works in this area include Synthetic Minority Oversampling TEchnique (SMOTE) [40], Borderline-SMOTE [51], Adaptive Synthetic Sampling [52], Safe-Level-SMOTE [53] and SPIDER2 [54] algorithms. Cost-sensitive learning considers higher costs for the misclassification of instances of the positive class (minority class) respecting the negative class (majority class), and therefore tries to minimize higher cost errors [42,43]. Thus, it is usually a solution that incorporates approaches at the data level, at the algorithmic level, or at both levels combined. For example, in the context of data level, the most popular method is resampling the training dataset according to the cost decision matrix by means of undersampling/oversampling [43] or assigning instance weights [55]. In the context of algorithmic level, the cost-sensitive algorithm based on the decision tree theory assigns the class label to the node that minimizes the classification cost [42,56]. Ensemble techniques construct several simple classifiers from the original data in order to aggregate their predictions when unknown instances are presented. To deal with the class imbalance problem, ensemble-based methods have brought along a growth of attention from researchers [44–46,57–59]. Also, ensemble techniques can be mainly categorized into two groups, i.e., cost-sensitive ensemble learning [57,60–62] and data preprocessing ensemble learning [44,58,63–66]. A complete taxonomy of ensemble techniques for learning with imbalanced classes has been proposed in well-known surveys [39,67].

CBC is relatively theoretically simple and computationally efficient compared with classifiers mentioned above. However, CBC is often plagued for inductive bias [10] or model misfit [9]. Thus, many researchers have proposed the improved methods of CBC, which can be mainly classified into three categories, i.e., choosing supervised term weighting, constructing good centroids in the initial phase, as well as adjusting the position of centroids during the training phase.

As pointed in [68], term weighting is a critical factor that affects the performance of the classifier. Therefore, term weighting methods for TC have gained increasing attention in many literatures. For instance, Lan et al. [69] proposed a new supervised Relevance Frequency (RF) factor for term weighting. Based on the RF factor, Nguyen et al. [70] proposed two robust factors, i.e., Kullback Leibler (KL) divergence and Jensen Shannon (JS) divergence, to overcome the problem that RF factor does not penalize terms