Neurocomputing 000 (2018) 1–19

Contents lists available at ScienceDirect

Neurocomputing

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

Exploiting efficient and effective lazy Semi-Bayesian strategies for text classification

Felipe Viegas[a,*], Leonardo Rocha[c], Elaine Resende[a], Thiago Salles[a], Wellington Martins[b], Mateus Ferreira e Freitas[b], Marcos André Gonçalves[a]

[a] Department of Computer Science, Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brazil
[b] Instituto de Informática, Universidade Federal de Goiás, Goiânia, GO, Brazil
[c] Department of Computer Science, Universidade Federal de São João Del Rei São João Del Rei, MG, Brazil

ARTICLE INFO

Article history:
Received 2 November 2017
Revised 26 February 2018
Accepted 12 April 2018
Available online xxx

Communicated by Dr. T. Mu

Keywords:
Text classification
Naive Bayes classifier
Semi-Naive Bayes heuristics
Feature weighting techniques
GPU parallelization

ABSTRACT

Automatic Document Classification (ADC) has become the basis of many important applications, e.g., authorship identification, opinion mining, spam filtering, content organizers, etc. Due to their simplicity, efficiency, absence of parameters, and effectiveness in several scenarios, Naive Bayes (NB) approaches are widely used as a classification paradigm. Due to some characteristics of real document collections, e.g., class imbalance and feature sparseness, NB solutions do not present competitive effectiveness in some ADC tasks when compared to other supervised learning strategies, e.g., SVMs. In this article, we investigate whether a proper combination of some alternative NB learning models with different feature weighting techniques is able to improve the NB effectiveness in ADC tasks and verify that comparable or even superior results when compared to the state-of-the-art in ADC can be achieved. Moreover, we also present an investigation on the relaxation of the NB attribute independence assumption (aka, Semi-Naive approaches) in large text collections, something missing in the literature. Given the high computational costs of these investigations, we take advantage of current many core GPU and multi-GPU architectures to perform such investigation, presenting a massively parallelized version of the NB approach. Finally, supported by the parallel implementations, we propose four novel Lazy Semi-NB approaches to overcome potential overfitting problems. In our experiments, the new lazy solutions are not only more efficient and effective than existing Semi-NB approaches, but also surpass, in terms of effectiveness, all other alternatives in the majority of the cases.

1. Introduction

The volume of data created and shared nowadays has reached unprecedented levels, making the organization and extraction of useful information from this huge amount of data one of the biggest challenges in Computer Science. Automatic Document Classification (ADC) has demonstrated to be a viable path towards this goal. Particularly, ADC techniques aim at building effective models capable of associating documents with well-defined semantic categories in an automated way. ADC techniques are the core component of many important applications, e.g., spam filtering [1], organization of topic directories [2], identification of writing styles or authorship [3], among many others.

ADC methods usually exploit a supervised learning paradigm [4], i.e., a classification model is first "learned" using previously labeled documents (training set), and then used to classify unseen documents (the test set). There are a plethora of supervised ADC algorithms available in the literature, e.g., Nearest-Neighbor classifiers [5], Support Vector Machines [6], boosting [7] and Bayesian models [8]. In this work, we focus on the Bayesian approach, due to its efficiency, simplicity, and effectiveness in several domains, especially in document classification. In particular, we focus on Naive Bayes methods, since they are is widely used for text classification.

Although a largely popular classification paradigm, other statistical learning methods, e.g., SVMs, have presented superior effectiveness when compared to Naive Bayes approaches in ADC. Naive Bayes is often used as a baseline in text classification because it is fast and easy to implement. However, some characteristics present in text collections, such as class imbalance and feature sparseness do compromise the Naive Bayes classification effectiveness [9,10].

* Corresponding author.
E-mail addresses: frviegas@dcc.ufmg.br (F. Viegas), lcrocha@ufsj.edu.br (L. Rocha), tsalles@dcc.ufmg.br (T. Salles), wellington@inf.ufg.br (W. Martins), mgoncalv@dcc.ufmg.br (M.A. Gonçalves).

https://doi.org/10.1016/j.neucom.2018.04.033
0925-2312/

Class imbalance happens when the number of documents on few classes is higher than that of most other classes in a dataset. This increases the bias of the trained classifier towards assigning most unseen documents to the largest classes, incurring in a poor classification effectiveness in the minority ones, the most important several applications (e.g. spam detection, vandalism).

Sparseness problem is related to the low frequency of certain features (i.e., terms/words) in some documents [11]. In Naive Bayes, the conditional probability of a term a_j given a class c_i is estimated using all training documents from c_i in which a_j occurs. Such conditional probabilities are negatively affected if a_j occurs only in a few documents, especially in smaller classes (i.e., those with few documents).

There are some proposals in the literature to overcome such problems, either by proposing some changes in the construction of the Naive Bayes learning model [9,12,13] or by means of feature weighting strategies more adequate to ADC, in a preprocessing phase prior to the model construction [10,11]. Although both research lines can produce significant improvements in the Naive Bayes effectiveness, when compared to its original version, the resulting methods are still not capable of surpassing some state-of-the-art ADC method, e.g., SVMs. Thus, as our first contribution we present a broad and original (never reported) study on the combination of different Naive Bayes-based learning models with different feature weighting strategies, evaluating these combinations in several real-world datasets. Our experimental results show that a proper combination of learning paradigms and weighting strategies produces results comparable and even superior to SVMs in several datasets, at a lower cost performance.

A third research line that has been investigated in order to improve the Naive Bayes effectiveness is the so-called *Semi-Naive Bayes* (Semi-NB) method, which relax the Naive Bayes attribute independence assumption [8], by reduction of data [14] or, mainly, by means of extensions of the structure of the learning model to represent feature dependencies [15,16]. These strategies have produced gains when compared to NB in small datasets, for instance, those related to bioinformatics [16]. However, due to their high computational cost, inherent to the complexity of representing the term dependencies, they cannot scale to large classification tasks. Thus, investigating whether the relaxation of the Naive Bayes attribute independence assumption is effective in large ADC tasks is still an open problem. In this context, we present our second and third contributions. The second contribution corresponds to an exclusive parallel, multi-GPU version of the Naive Bayes approach using graphic processing units (GPUs). This parallel version allowed us to implement some Semi-NB approaches, capable of running in large text collections. Thus, our third contribution is an original study about the impacts of the relaxation of the Naive Bayes assumption in large ADC tasks.

Finally, our fourth contribution is the introduction of four original parallel lazy Semi-Naive Bayes strategy proposals that exploit the information of the document to be classified (i.e., lazy) to reduce the complexity of the Semi-Naive Bayes learned the model.We hypothesize that that Semi-NB solutions tend to loose performance due to overfitting issues caused by the strong dependencies found on the validation set. Accordingly, we propose to mitigate this problem by proposing novel Lazy Semi-NB. These solutions only look for dependencies in the context of the terms that occur in the (test) document to be classified. Our experimental results point out that further improvements can be obtained with these models and that our final solutions, with their improvements, are able to obtain results that are better than some state-of-the-art methods in ADC.

In summary, the main research questions we address in this work are: (Q1) Can a proper combination of a Naive Bayes learning paradigm with feature weighting strategies, specially designed to deal with the ADC idiosyncrasies, be competitive or surpass state-of-the-art classifiers? (Q2) Can we design an efficient (massively parallel) Naive Bayes implementation that allows testing interesting, but very costly, proposals that relax the Naive Bayes independence assumption in large ADC tasks? (Q3) If yes for (Q2), are these Semi-Naive Bayes proposals capable of improving even further the best combinations found in the answer of (Q1)? and (Q4) Are lazy extensions of these proposals capable of overcoming the overfitting problem commonly found in Semi-NB solutions?

The remainder of this work is organized as follows. Section 2 covers related work with focus on extensions of of the NB model that relax the independence assumptions and the use of GPU-based parallelism in Machine Learning. In Section 3, we describe the extensive combinations of feature weighting strategies with NB models we exploit. We also describe the main Semi-Naive Bayes models in the literature. In Section 4, we introduce our main Semi-NB approach, called Lazy Super Parent Tree Augmented Naive Bayes (LSPTAN) and its variations, especially designed for ADC. Section 5 presents our extensive experimental evaluation. In Section 6, we describe our GPU-based implementation of Bayesian Classifiers. We first introduce our GPU-based parallel implementation of the Naive Bayes algorithm and then we describe our GPU-based parallel implementation of the LSPTAN strategies. In Section 7, we quantify the factors that explain our effectiveness gains and the speedups in efficiency of our solutions. We conclude the article with future work in Section 8.

2. Literature review

In this section, we present the strategies proposed in the literature to deal with the problems discussed in the previous section. First, we discuss some aspects that limit the effective the Naive Bayes models. Then, we present some strategies that perform some adjustments in the Naive Bayes model, making it more robust for text classification. Next, we present some proposals in the literature that extend the Bayesian model, alleviating the assumption of independence between attributes assumed by the Naive Bayes algorithm: the so-called Semi-Naive Bayes strategies. Finally, we discuss the use of GPU-based parallel implementations in classification algorithms.

2.1. Naive Bayes limitations

A fundamental assumption assumed by most automatic classifiers is that the data used to learn the classification model are random samples independently and identically distributed (i.i.d.) from a statistical distribution that governs the data. However, this may not be the case. Indeed, in many real scenarios, the training data do not follow the same distribution of the test data, compromising the effectiveness of the classification algorithms.

The Naive Bayes (NB) algorithm is one of the most widely used techniques, due to its simplicity and efficiency in several scenarios, especially when applied to datasets in which the attributes are independent, making their “Naive” assumption more reliable. Normally in text classification, its effectiveness is not as good as some other statistical learning methods. This is due to some characteristics presented in real document collections, such as skewness (aka, class imbalance) and sparsity that compromise some of Naive Bayes premises. In [9,10], the authors described some problems faced by the Naive Bayes classifiers when applied to real textual scenarios. The principal factors are *class imbalance, document length* and *feature sparseness*.

Skewness [17] refers to scenarios in which the number of documents of one or few classes far exceeds the number of documents in the other classes. A number of documents in classes is an important factor since it is related to the amount of information used to

Please cite this article as: F. Viegas et al., Exploiting efficient and effective lazy Semi-Bayesian strategies for text classification, Neurocomputing (2018), https://doi.org/10.1016/j.neucom.2018.04.033