Knowledge-Based Systems 85 (2015) 71–79

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

A gradient approach for value weighted classification learning in naive Bayes

Chang-Hwan Lee

Department of Information & Communications, Dongguk University, Seoul, Republic of Korea

ARTICLE INFO

Article history:
Received 3 October 2014
Received in revised form 6 April 2015
Accepted 18 April 2015
Available online 29 April 2015

Keywords:
Classification
Bayesian learning
Feature weighting
Gradient descent

ABSTRACT

Feature weighting has been an important topic in classification learning algorithms. In this paper, we propose a new paradigm of assigning weights in classification learning, called value weighting method. While the current weighting methods assign a weight to each feature, we assign a different weight to the values of each feature. The proposed method is implemented in the context of naive Bayesian learning, and optimal weights of feature values are calculated using a gradient approach. The performance of naive Bayes learning with value weighting method is compared with that of other state-of-the-art methods for a number of datasets. The experimental results show that the value weighting method could improve the performance of naive Bayes significantly.

1. Introduction

In some classifiers, the algorithms operate under the implicit assumption that all features are of equal value as far as the classification problem is concerned. However, when irrelevant and noisy features influence the learning task to the same degree as highly relevant features, the accuracy of the model is likely to deteriorate. Since the assumption that all features are equally important hardly holds true in real world application, there have been some attempts to relax this assumption in classification. Zheng and Webb [22] provide a comprehensive overview of work in this area.

The first approach for relaxing this assumption is to combine feature subset selection with classification learning. It is to combine a learning method with a preprocessing step that eliminates redundant features from the data.

Feature selection methods usually adopt a heuristic search in the space of feature subsets. Since the number of distinct feature subsets grows exponentially, it is not reasonable to do an exhaustive search to find optimal feature subsets. In the literature, it is known that the predictive accuracy of naive Bayes can be improved by removing redundant or highly correlated features [13]. This makes sense as these features violate the assumption that each feature is independent on each other.

Another major way to help mitigate this weakness, feature independence assumption, is to assign different weights to different features in classification learning. Since features do not play the same role in many real world applications, some of them are more important than others. Therefore, a natural way to extend classification learning is to assign each feature different weight to relax the conditional independence assumption. Feature weighting is a technique used to approximate the optimal degree of influence of individual features using a training set. While feature selection methods assign 0/1 values as the weights of features, feature weighting is more flexible than feature subset selection by assigning continuous weights. Therefore, feature selection can be regarded as a special case of feature weighting where the weight value is restricted to have only 0 or 1.

Even though there have been many feature weighting methods proposed in the literature, many of them have been applied in the domain of nearest neighbor algorithms [20], and have significantly improved the performance of nearest neighbor methods.

In this paper, we propose a new paradigm of weighting method, called *value weighting* method. While the current weighting methods assign a weight to each *feature*, we assign a weight to each *feature value*. Therefore, the value weighting method is a more fine-grained weighting method than the feature weighting method. While there have been a few work focused on feature weighting in the literature, to our best knowledge, there has been no work which assigns different weight to each feature value in classification leaning. We extended the current hypothesis space of classification learning into next level by introducing a new set of weight space to the problem. Therefore, this paper proposes a new dimension of weighting method by assigning weights to feature values.

The new value weighting method provides a potential of expanding the expressive power of classification learning, and possibly improves its performance. Since features weighting

E-mail address: chlee@dgu.ac.kr

http://dx.doi.org/10.1016/j.knosys.2015.04.020
0950-7051/

methods improve the performance of the classification learning, we will investigate whether assigning weights to feature *values* can improve the performance even further.

The main contribution of this paper is to provide a new paradigm of weighting method in classification learning. In this paper, we study the value weighting method in the context of naive Bayesian algorithm. We have chosen naive Bayesian algorithm as the template algorithm since it is one of the most common classification algorithms, and many researchers have studied the theoretical and empirical results of this approach. It has been widely used in many data mining applications, and performs surprisingly well on many applications [5].

There have been only a few methods for combining feature weighting with naive Bayesian learning [10,11,21]. The feature weighting methods in naive Bayes are known to improve the performance of classification learning. We will investigate, in this paper, whether the value weighting method provides enhanced performance in the context of naive Bayes.

The basic assumption behind the value weighting method in this paper is that each feature value has different significance with respect to class value. When we say a certain feature is important/significant, we believe that the importance of the feature can be decomposed. For instance, suppose we are to learn about a rare form of pregnancy-related disease, and try to calculate the importance of feature *gender*. Obviously, if the value of gender is *male*, that observation has no effect with respect to the target feature(pregnancy-related disease). On the other hand, the value of *female* in gender feature surely has significant impact to the target feature. If we assign the same weight to each feature value, we will lack the capability to discriminate the predicting power residing across feature values.

The rest of this paper is structured as follows. In Section 2, we describe the related work on weighting methods in naive Bayesian learning. Section 3 shows the basic concepts of value weighted naive Bayesian learning, and Section 4 discusses the mechanisms of the new value weighting method. Section 5 shows the experimental results of the proposed method, and Section 6 summarizes the contributions made in this paper.

2. Related work

A number of approaches have been proposed in the literature of feature selection [19,13,14,9]. While feature selection methods assign 0/1 value to each feature, feature weighting assigns a continuous value weight to each feature. While there have been many research for assigning feature weights in the context of nearest neighbor algorithms [20], very little work of weighting features is done in naive Bayesian learning.

Hall [11] proposed a feature weighting algorithm for naive Bayes using decision tree. The method is called DTNB, and it estimates the degree of feature dependency by generating full decision trees and looking at the depth at which features are tested in the decision tree. They show that DTNB could improve the performance of traditional naive Bayes.

Lee et al. [15,16] proposed an information-theoretic method for calculating feature weights in naive Bayes. They calculated the feature weights of naive Bayes using Kullback–Leibler measure. The averaged amount of information for a feature is calculated, and it showed improvement over normal naive Bayes and other supervised learning methods.

Cardie [3] used an information gain based on the position of a feature in a decision tree to derive feature weights for nearest neighbor algorithm. They designed case-based learning algorithms to improve the performance of minority class predictions. They used local weighting method, and weights are derived for each test instance based on their path it takes in the tree.

Gartner [10] employs SVM algorithm for feature weighting. The algorithm looks for an optimal hyperplane that separates two classes in given space. The weights associated with the hyperplane can be regarded as feature weights in the naive Bayes. It can solve the binary classification problems and the feature weight is based on conditional independence. They showed that the method shows better performance than other state-of-the-art machine learning methods.

3. Background

In this paper, we propose a new value weighting method in the context of naive Bayesian learning. The naive Bayesian classifier is a straightforward and widely used method for supervised learning. It is one of the fastest learning algorithms, and can deal with any number of features or classes. Despite of its simplicity in model, naive Bayes performs surprisingly well in a variety of problems. Furthermore, naive Bayesian learning is robust enough that small amount of noise does not perturb the results.

Two fundamentally different approaches to this optimization problem can be identified, the filter-based and the wrapper-based. The class of filter-based methods contains algorithms that use no input other than the training data itself to calculate the feature weights, whereas wrapper-based algorithms use feedback from a classifier to guide the search. Wrapper-based algorithms are inherently more powerful than their filter-based counterpart as they implicitly take the inductive bias of the classifier into account. In this paper we implement the value weighting method in the context of naive Bayes using wrapper method.

The naive Bayesian learning uses Bayes theorem to calculate the most likely class label of the new instance. Since all features are considered to be independent given the class value, the classification on d is defined as follows

$$\mathcal{V}_{NB}(d) = \text{argmax}_c P(c) \prod_{a_{ij} \in d} P(a_{ij}|c)$$

where a_{ij} represents the j-th value of the i-th feature. In spite of its good performance, naive Bayesian classifier is known to be a linear classifier, and thus has some difficulties with solving non-linearly separable problems.

The naive Bayesian classification with feature weighting is now represented as follows.

$$\mathcal{V}_{FWNB}(d) = \text{argmax}_c P(c) \prod_{a_{ij} \in d} P(a_{ij}|c)^{w_i} \quad (1)$$

where $w_i \in \mathbb{R}$ represents the weight of feature. In this formula, unlike traditional naive Bayesian approach, each feature i has its own weight w_i. Feature weighting is a generalization of feature selection, and involves a much larger search space than feature selection.

The *feature weighted* naive Bayesian classifier provides weight to each feature, and could improve the performance of regular naive Bayesian learning, However, as shown in Lemma 1, the *feature weighted* naive Bayesian classifier still belongs to the category of linear classifier. For simplicity, suppose an instance d has n binary features, such as $d = (a_1, \ldots, a_n)$, and there are only two classes, $c = 1$ and $c = 0$. From the definition of linear classifier, a classifier is linear if it can be represented in a form of $\mathcal{V} = \text{sign}(\sum_i w_i a_i)$ (assuming $a_0 = 1$) where w_i is the constant coefficient associated with a_i.

Lemma 1. *Suppose*

$$s_i = \log \frac{p(a_i = 1|c = 1)}{p(a_i = 1|c = 0)} \quad \text{and} \quad t_i = \log \frac{p(a_i = 0|c = 1)}{p(a_i = 0|c = 0)}$$