Neurocomputing 143 (2014) 248–260

Contents lists available at ScienceDirect

Neurocomputing

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

Impact of a metric of association between two variables on performance of filters for binary data

Kashif Javed [a,*], Haroon A. Babri [a], Mehreen Saeed [b]

[a] Department of Electrical Engineering, University of Engineering and Technology, Lahore, Pakistan
[b] Department of Computer Science, National University of Computer and Emerging Sciences, Lahore, Pakistan

ARTICLE INFO

Article history:
Received 9 March 2014
Received in revised form
1 May 2014
Accepted 31 May 2014
Communicated by Feiping Nie
Available online 11 June 2014

Keywords:
Feature selection
Filters
Binary data
Classification

ABSTRACT

In the feature selection community, filters are quite popular. Design of a filter depends on two parameters, namely the objective function and the metric it employs for estimating the feature-to-class (relevance) and feature-to-feature (redundancy) association. Filter designers pay relatively more attention towards the objective function. But a poor metric can overshadow the goodness of an objective function. The metrics that have been proposed in the literature estimate the relevance and redundancy differently, thus raising the question: can the metric estimating the association between two variables improve the feature selection capability of a given objective function or in other words a filter. This paper investigates this question. Mutual information is the metric proposed for measuring the relevance and redundancy between the features for the mRMR filter [1] while the MBF filter [2] employs correlation coefficient. Symmetrical uncertainty, a variant of mutual information, is used by the fast correlation-based filter (FCBF) [3]. We carry out experiments on mRMR, MBF and FCBF filters with three different metrics (mutual information, correlation coefficient and diff-criterion) using three binary data sets and four widely used classifiers. We find that MBF's performance is much better if it uses diff-criterion rather than correlation coefficient while mRMR with diff-criterion demonstrates performance better or comparable to mRMR with mutual information. For the FCBF filter, the diff-criterion also exhibits results much better than mutual information.

1. Introduction

In data and knowledge management systems, feature selection is frequently used as a preprocessing step for data analysis [4,5]. The aim is to remove from the data irrelevant and redundant information [6]. To formally define the feature selection problem, we first introduce some notation. Suppose that $\mathcal{D}=\{\mathbf{x}^t, C^t\}_{t=1}^{N}$ represents a labeled data set consisting of N instances and M features such that $\mathbf{x}^t \in R^M$ and C^t denotes the class variable of instance t. In this study, we assume that variable C can attain two values, thus representing a two-class classification task. Each vector $\mathbf{x}$ is, thus, an M-dimensional vector of features, i.e., $\mathbf{x}^t=\{F_1^t, F_2^t, \ldots, F_M^t\}$. Let $W_k(F_i, C)$ denote the relationship estimated by the kth metric of association, where the two variables are the ith feature and C. Furthermore, $\mathcal{F}$ is used to refer to the set comprising all features of a data set, whereas $\mathbf{G}$ denotes a feature subset. The feature selection problem is to find a subset $\mathbf{G}$, termed as the final or optimal subset, of m features from the set $\mathcal{F}$ having M features with the smallest classification error [7] or at least without a significant degradation in the performance [8].

Removal of irrelevant and redundant features through a feature selection algorithm improves the effectiveness and efficiency of learning algorithms [2,9]. Additionally, the learned results become comprehensible [3]. Because of these advantages, feature selection has been a fertile field of research and development over the years. Different approaches have been suggested and are broadly categorized as filters, wrappers, and embedded methods [10]. Filters select a subset of most useful features as a standalone task independent of the learning algorithm [11]. Wrappers search for a good subset in the feature space with the guidance of a classifier [7]. In the embedded approach, the feature selection process is integrated into the training process of a given classifier [12]. Among these methods, filters are highly popular because of their fast speed of computation and less chances of overfitting [13,14]. Recently, researchers have shown interest in FS methods based on sparse learning because of their good performance [15]. For example, the method of Nie et al. [16] was found to outperform the well-known minimal-redundancy-maximal-relevance (mRMR) and information gain (IG) algorithms. In this study, our focus is however only on filter methods.

As filters select features independently of any particular learning algorithm, these methods require a function or a criterion to

* Corresponding author.
E-mail addresses: kashif.javed@uet.edu.pk (K. Javed), babri@uet.edu.pk (H.A. Babri), mehreen.saeed@nu.edu.pk (M. Saeed).

http://dx.doi.org/10.1016/j.neucom.2014.05.066
0925-2312/

quantify how useful is the presence of a feature in the final subset. The function is optimized so that features highly relevant to the class but having minimum redundancy get selected. To meet this goal, a metric of association is used to estimate the relevance between the feature and the class variable and the redundancy of features. Although the metric is equally important for a filter, filter designers are usually more concerned about the working of the objective function. The usefulness of a filter (or an objective function) is established by testing it against other filters. This however only provides insight into the performance of a filter with a given metric.

A number of metrics have been proposed in the literature [17]. These metrics of association measure the relationship of two variables differently [18,19]. A metric that poorly estimates the relevance and redundancy of features can spoil the goodness of a filter's objective function. On the other hand, a filter that has shown poor performance against other filters can demonstrate better performance if a better metric of association is chosen for it. This aspect of a filter's working has been ignored. In this work, we analyze filters from this perspective and investigate how the performance of a given filter is affected when it uses different metrics meant for measuring the relationship between variables. For this purpose, we examine the performance of three well-known filters, namely minimal-redundancy-maximal-relevance (mRMR) filter [1], Markov blanket filter (MBF) [2] and fast correlation-based filter (FCBF) [3] with three different metrics (mutual information, correlation coefficient and diff-criterion). Peng et al. [1] suggested that mutual information should be employed for the mRMR filter. We use this combination as a reference to see whether the filter's performance improves when it uses correlation coefficient and diff-criterion. Similarly, the MBF and correlation coefficient combination as suggested by Koller and Sahami [2] can be taken as a reference to investigate the performance of MBF with mutual information and the diff-criterion. For the FCBF, Yu and Liu [3] employed symmetrical uncertainty (a variant of mutual information). We consider FCBF and mutual information to be the reference against which the other combinations will be tested. Because of their wide applications [20–25], experiments are carried out on binary data sets. We use three binary data sets from different application domains with four widely used classifiers to evaluate the performance of filters.

The remainder of the paper is organized into four sections. Section 2 describes the theory related to filters and discusses well-known filter methods. We also present various metrics of association. In Section 3, experimental setup is described while results obtained on three real-life data sets are presented and discussed in Section 4. The conclusions are drawn in Section 5.

2. Filters

The idea of filters is to separate feature selection from classification. Filters search the feature space by optimizing a function or a criterion, which is termed as the objective function. The goal is to select those features in the final subset that maximize the relevance and minimize the redundancy. The function therefore acts as a proxy measure of the accuracy of the classification algorithm [14]. It employs a metric of association, which estimates the association between variables by taking the statistics of the data into account. Because of these characteristics, the filter's search for an optimal feature subset is less expensive than that of wrappers and embedded methods [26]. Filters are also known to be less prone to overfit the training data [13]. These advantages have made filters to be widely used among researchers of different application domains, which are also highly active in designing new filters.

If the design of filters is taken into consideration, we can broadly identify two design parameters: an objective function with which a filter populates the final subset by searching for the most useful features and a metric of association used for estimating the usefulness of features. The combination of these two components determines the performance of a filter. However, one can find that filters that have been proposed in the literature [27,28,14] primarily vary in the objective function. Even though a number of metrics of association have been proposed in the literature, metrics have drawn less attention of the filter designers. Because of this, different filters may even employ the same metric without considering its impact on the objective function. For instance, Battiti's mutual information based feature selection (MIFS) criterion [29] and Peng et al.'s mRMR filter [1] both use mutual information as the metric of association but optimize different objective functions.

Based on a metric that is not a good estimator of the relevance and redundancy, even a well-designed objective function will not be able to select the most useful features thus, resulting in poor performance. One can find many studies in the literature [11,30] that compare the performance of different filters. But hardly any study exists which investigates the effect of the metrics of association on the performance of a given objective function. The selection of a good metric is therefore important for a filter. The motivation behind this argument is illustrated with an example given in Section 2.3. But before that, let us next present those metrics of association and filters that are used for this study.

2.1. Metrics of association

Because of computational issues, filters generally consider pairwise interactions among variables, i.e., feature-to-class and feature-to-feature relationships [31]. A number of metrics have been proposed in the literature and can be broadly categorized into three groups, namely correlation based, information-theoretic based and probabilistic metrics [32,17]. Although the expressions given in this section are written in terms of a feature (F_i) and the class variable (C), we can also use the same expression for estimating the redundancy between two features by replacing C with a feature, say F_j.

2.1.1. Correlation based metrics

Pearson's correlation coefficient [33] and Chi-squared statistics [34] are well-known members of this category. The correlation coefficient (CC) is designed to estimate the linear relationship between two variables and is given by

$$W_{cc}(F_i, C) = \frac{E(F_i C) - E(F_i)E(C)}{\sqrt{\sigma^2(F_i)\sigma^2(C)}} \tag{1}$$

where E represents the expected value of a variable(s) and σ^2 denotes its variance. When F_i and C are linearly dependent, $W_{cc}(F_i, C)$ is equal to ± 1 and when they are completely uncorrelated, $W_{cc}(F_i, C)$ becomes 0.

2.1.2. Information-theoretic metrics

In this category, mutual information [29] and symmetrical uncertainty [3] are widely known. Mutual information (MI) estimates the relationship between the joint distribution $p(C, F_i)$ of two variables C and F_i and their product distribution $p(C)p(F_i)$ [35] and is given by

$$W_{mi}(F_i, C) = \sum_{C}\sum_{F_i} p(C, F_i) \log \frac{p(C, F_i)}{p(C)p(F_i)} \tag{2}$$

The value of $W_{mi}(F_i, C) \geq 0$ and a larger value means that a variable contains more information about the other variable. When two variables have no information in common, $W_{mi}(F_i, C)$ reduces to 0.