Knowledge-Based Systems 71 (2014) 322–338

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Feature selection by multi-objective optimisation: Application to network anomaly detection by hierarchical self-organising maps

Emiro de la Hoz [a], Eduardo de la Hoz [a], Andrés Ortiz [b,*], Julio Ortega [c], Antonio Martínez-Álvarez [d]

[a] Programa de Ingeniería de Sistemas, Universidad de la Costa, Barranquilla, Colombia
[b] Communications Engineering Department, University of Málaga, Málaga, Spain
[c] Computer Architecture and Technology Department, CITIC, University of Granada, Spain
[d] Computer Technology Department, University of Alicante, Alicante, Spain

ARTICLE INFO

Article history:
Received 22 February 2014
Received in revised form 15 July 2014
Accepted 11 August 2014
Available online 20 August 2014

Keywords:
Feature selection
Multi-objective optimisation
Unsupervised clustering
Growing self-organising maps
Network anomaly detection
IDS

ABSTRACT

Feature selection is an important and active issue in clustering and classification problems. By choosing an adequate feature subset, a dataset dimensionality reduction is allowed, thus contributing to decreasing the classification computational complexity, and to improving the classifier performance by avoiding redundant or irrelevant features. Although feature selection can be formally defined as an optimisation problem with only one objective, that is, the classification accuracy obtained by using the selected feature subset, in recent years, some multi-objective approaches to this problem have been proposed. These either select features that not only improve the classification accuracy, but also the generalisation capability in case of supervised classifiers, or counterbalance the bias toward lower or higher numbers of features that present some methods used to validate the clustering/classification in case of unsupervised classifiers.

The main contribution of this paper is a multi-objective approach for feature selection and its application to an unsupervised clustering procedure based on Growing Hierarchical Self-Organising Maps (GHSOMs) that includes a new method for unit labelling and efficient determination of the winning unit. In the network anomaly detection problem here considered, this multi-objective approach makes it possible not only to differentiate between normal and anomalous traffic but also among different anomalies. The efficiency of our proposals has been evaluated by using the well-known DARPA/NSL-KDD datasets that contain extracted features and labelled attacks from around 2 million connections. The selected feature sets computed in our experiments provide detection rates up to 99.8% with normal traffic and up to 99.6% with anomalous traffic, as well as accuracy values up to 99.12%.

1. Introduction

Frequently, existing classification problem features are not discriminative enough. Moreover, the use of correct features improves classification performance and reduces computational time. Thus, feature extraction and selection are two important classification problem issues that aim to obtain a subset of features in a lower dimensional space. They provide different advantages:

1. Representing the data in a lower dimensional space avoids the *curse of dimensionality* [1,2]:
   - Diminishes the number of examples needed to train a classifier. The number of train samples grows exponentially with the data dimensionality.
   - Avoids overfitting and improves the classifiers' generalisation performance. In practice, there is an optimum number of features for maximum classification performance.
2. High informative features will represent the different class samples far away in the feature space, while similar samples will be represented close to each other.
3. The use of fewer features improves computational efficiency.
4. Data visualisation is easier and more intuitive in lower dimensional spaces (for example, 2D or 3D).

However, feature extraction and feature selection are different processes, although both can be used to obtain a discriminative subset. Thus, we will describe both separately.

* Corresponding author.
E-mail addresses: edelahoz@cuc.edu.co (E. de la Hoz), edelahoz6@cuc.edu.co (E. de la Hoz), aortiz@ic.uma.es (A. Ortiz), jortega@ugr.es (J. Ortega), amartinez@dtic.ua.es (A. Martínez-Álvarez).

http://dx.doi.org/10.1016/j.knosys.2014.08.013
0950-7051/

Feature extraction can be stated as follows: Let $x \in \mathbb{R}^n$ be the existing feature set. The goal is to make existing features more descriptive through a mapping function $g(x) : \mathbb{R}^n \rightarrow \mathbb{R}^m$ in such a way that $\hat{x} = g(x)$ preserves the information and structure of data in $\mathbb{R}^n$. Thus, in general, $g(x)$ may implement a non-linear mapping. Frequently, linear transformations through a matrix $\mathbb{H}$, are usually applied to revise the initial feature set $x\ \hat{x} = \mathbb{H}^T x$. In this case, $\hat{x}$ is the representation of x in the subspace spanned by the basis vectors in $\mathbb{H}$ [3].

A representative feature extraction example through linear mapping is Principal Component Analysis (PCA). PCA generates an orthonormal basis vector indicating the maximum variance directions. Thus, projecting onto this basis maximises the sample scatter. The data samples projected onto the low dimensional space spanned by the central components are used in the classification task. Another popular feature extraction technique that uses a classification criterion instead of the representation error (as in PCA), is Linear Discriminant Analysis (LDA) [1]. In this case, the samples may not be accurately represented by the projected features (that is, reconstruction error is not minimised), but class discriminative information is enhanced. PCA and LDA have been used in classical problems, such as facial recognition, *eigenfaces* [4] and *fisherfaces* [5], respectively. Other techniques such as *Independent Component Analysis* (ICA), [6] aim to find a linear representation of non-Gaussian data in such a way that the components are as statistically independent as possible.

Unlike extraction, selection does not transform the existing features, but only searches for the most informative subset. Feature selection algorithms are classified into two categories: *filters*, and *wrappers*. *Filters* do not use a classifier and evaluate the features according to heuristics that accommodate different data characteristics. Thus, features are ranked according to their importance for separating classes using either statistical methods, information theory-based methods or searching techniques. Statistical methods include hypothesis testing, such as the Student's *t-test* [7,1]. Other statistical methods, such as the Fisher Discriminant Ratio, can be used to quantify the discriminative power of individual features between two equiprobable classes [1]. Information theory-based methods can use different metrics, such as Entropy, Kullback–Leibler divergence [1] or the information gain measure [8] to rank the features. Moreover, [9,10] use the Conditional Mutual Information (CMI) as the criterion for selecting feature subsets. Other *filter* algorithms use a correlation-based metric to evaluate feature usefulness. Specifically, the Correlation-Based Feature Selection (CFS) algorithm [11,12] takes into account individual feature worth based on the hypothesis that *good feature subsets contain features highly correlated with the class, yet uncorrelated with each other* [11].

Nevertheless, most *filter* methods evaluate feature usefulness for predicting class labels by computing an average score on the different dataset classes. This may lead to removing features from the final selection that could be specially relevant for a certain class label. Thus, it is necessary to evaluate the discriminative power of each feature, selecting those that best describe each individual class. This is especially useful for imbalanced datasets or data with different statistical class distribution. In order to overcome this limitation, [13] proposed a framework focused on evaluating feature relevance and redundancy to a certain class label.

Unlike *filters*, *wrappers* use an objective function that returns the current feature selection goodness. This feedback is obtained from the classifier outcome (that is, classification accuracy or classification error) executed on the training set. However, these approaches are classifier-dependent, and require executing the training process in each iteration. Different searching strategies can be used depending on the way the features are selected or discarded in each iteration. However, their common goal is to keep the best feature combination (that is, features that optimise the objective function). In this way, there are two main searching strategies:

1. Suboptimal searching. These techniques aim to avoid trying all the feature combinations. Sequential Forward Selection (SFS) and Sequential Backward Selection (SBS) are well-known suboptimal searching methods [1].
2. Exhaustive searching. All possible feature combinations will be used to train the classifier and to assess performance. It is computationally expensive and may be unfeasible for high-dimensionality feature spaces or large datasets.

In *wrapper* feature selection approaches, evaluating a given set's utility presents different issues depending on the type, supervised or unsupervised, of the classification procedure used. If the procedure is supervised, it is relatively easy to define the utility cost function by using the classification error. Nevertheless, in unsupervised procedures, the utility should be determined from a clustering quality definition without having knowledge about the corresponding labels or even the number of clusters. Frequently, the clustering quality measures use ratios between intra-cluster compactness measures and inter-cluster separation ones. Nevertheless, the distances between points tend to be similar values as the dimensions are higher, making these quality solutions biased toward lower dimension solutions [2]. This way, although, as is indicated in [2], formulating feature selection as a multi-objective optimisation problem could provide some advantages, results would depend on whether the procedure is either supervised or unsupervised. In the supervised classification procedures, the goal is usually maximising the classifier performance while the number of features is minimised as larger sets could produce overfitting and low generalisation problems. This way, a multi-objective optimisation approach that takes into account the classifier performance and the number of features allows for an adequate formulation of this goal. The situation in unsupervised classification problems is different. In this case, it is difficult to evaluate the clustering and, as has been previously indicated, the applied validation techniques usually present a dimensionality bias to either smaller or larger cardinality feature sets. Thus, a multi-objective approach could counterbalance the specific bias of the considered cluster validation method. Here, we propose using the NSGA-II algorithm [14] for multi-objective optimisation to build a wrapper approach that selects specific feature subsets for each class label. Some other works have been proposed to implement feature selection as a multi-objective optimisation, either for supervised or unsupervised classifiers. They are referenced and compared with the approach here proposed in Section 5.

In this paper, feature selection is considered in the context of network intrusion detection systems. With the growth of Internet, not only the number of interconnected computers, but also the relevance of network applications, have increased considerably. At the same time, the trend to online services has exposed a lot of sensitive information [15,16]. This way, there are three main alternatives for protecting information. The first consists of avoiding sending information in clear (without any encryption). Such systems encrypt the information before sending for keeping its privacy. The second consists of using a separate physical or logical channel to transfer the information. This is the case of the Virtual Private Networks (VPN), which emulate a dedicated connection between two hosts. As a third alternative, the information on the VPNs can also be encrypted. Nevertheless, there is not any infallible encryption method and the encryption/decryption process can