Feature Selection By Multi-Objective Optimisation: Application To Network Anomaly Detection By Hierarchical Self-Organising Maps

Emiro De La Hoz Franco; Eduardo Miguel De La Hoz Correa; Ortiz, Andrés; Ortega, Julio; Martínez-Álvarez, Antonio

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%.

Keywords

Feature Selection; Growing Self-Organising Maps; IDS; Multi-Objective Optimization; Network Anomaly Detection; Unsupervised Clustering.