Review on Feature Selection and Classification using Neuro-Fuzzy Approaches

Saroj Biswas, Computer Science and Engineering Department, National Institute of Technology Silchar, Silchar, India
Monali Bordoloi, Computer Science and Engineering Department, National Institute of Technology Silchar, Silchar, India
Biswaajit Purkayastha, National Institute of Technology Silchar, Silchar, India

ABSTRACT

This research article attempts to provide a recent survey on neuro-fuzzy approaches for feature selection and classification. Feature selection acts as a catalyst in reducing computation time and dimensionality, enhancing prediction performance or accuracy and curtailing irrelevant or redundant data. The neuro-fuzzy approach is used for feature selection and for providing some insight to the user about the symbolic knowledge embedded within the network. The neuro–fuzzy approach combines the merits of neural network and fuzzy logic to solve many complex machine learning problems. The objective of this article is to provide a generic introduction and a recent survey to neuro-fuzzy approaches for feature selection and classification in a wide area of machine learning problems. Some of the existing neuro-fuzzy models are also applied on standard datasets to demonstrate the applicability of neuro-fuzzy approaches.

KEYWORDS

Artificial Neural Networks, Data Mining, Feature Selection, Fuzzy Logic, Neuro-Fuzzy System

1. INTRODUCTION

The focus of this era is not simply serving the purpose of a work but to optimize the process involved, in order to minimize time and space complexity. Machine learning algorithms in pattern recognition, image processing and data mining mainly ensure classification. These algorithms operate on a huge amount of data with multiple dimensions, from which knowledge is extracted. However, the entire dataset in hand does not always prove to be significant to each and every domain. An important concept that contributes extensively in classification and better understanding of the domain is feature selection (Kohavi and John, 1997). Feature selection is a process of selecting a subset of features from a set of features in a balanced manner, without losing most of the characteristics and identity of the original object. There are two factors that affect feature selection – irrelevant features and redundant features (Dash and Liu, 1997). Irrelevant features are those which provide no useful information in that context and redundant features are those which provide the same information as the currently selected features.

Selection of an optimal number of distinct features contributes substantially in the improvement of the performance of a classification system with lower computational effort, data visualization and improved understanding of computational models. Feature selection also reduces running time of learning algorithm, risk of data over fitting, dimensions of the problem and cost of future data.
acquisition (Guyon and Elisseeff, 2003). Thus, in order to cope up with the rapidly evolving data, many researchers have been proposing different feature selection techniques for classification tasks.

The main goals of feature selection are to select the smallest feature subset that yields the minimum generalization error, to reduce time complexity and to reduce memory and money for handling large datasets (Vergara and Estévez, 2014). In most common scenarios, feature selection methods are used for solving classification problems or are a part of a classification problem. Many classical techniques exist for the purpose of feature selection such as Mutual Information (MI), decision tree, Bayesian network, genetic algorithm, Support Vector Machine (SVM), K-nearest neighbor (K-nn), Pearson correlation criteria, Linear Discriminant analysis (LDA), Artificial Neural Network (ANN), Fuzzy sets. The choice of using a specific algorithm is a critical step as no such best algorithm exists that fits for considering every scope and solving every problem of feature selection and classification.

The use of Mutual Information (MI) for feature selection can be found in many contributions by different researchers (Vergara and Estévez, 2014; Peng et al., 2005; Grande et al., 2007; Chandrasekhar and Sahin, 2014; Battiti, 1994). Mutual information provides the dependencies between variables in terms of their probabilistic density functions. However, if one among the two variables is continuous, a limited number of samples obtained after feature selection makes the computation of the integral in the continuous space a bit challenging. (Peng et al., 2005). It has also been found that MI does not work efficiently in high-dimensional spaces and there exists no standard theory for MI normalization (Vergara and Estévez, 2014).

Decision tree is a tree based knowledge representation methodology used to represent classification rule; where at each node one can select the most useful feature for classification using some appropriate estimation criteria. Decision tree methods like ID3 and CART, in general, split a node using only a single attribute. However, use of combinations of attributes makes more sense in some cases (Setiono and Liu, 1997). Also, their performance degrades if they face many features which are unnecessary for the desired output prediction (Kohavi and John, 1997). In decision tree interpretation, the constructed tree may contain sub-tree(s) containing irrelevant attributes even when the data is not noisy and which misleads the user’s interpretation and thus lead to over-fitting and even data fragmentation.

Bayesian networks are directed acyclic graphs representing the joint probability distribution over a set of random variables (features). However, it suffers from the limitation that the number of structure super-exponentially increases as number of features increase and focuses more on the dependency of the features rather than the important features. In Naïve Bayes classifier, classification is done using posterior probability distribution by applying Bayes rule, but all attributes are assumed to be conditionally independent. Even if the relevant features are considered, the performance may degrade quickly if correlated features are added (Kohavi and John, 1997).

In genetic algorithm (GA), the chromosome bits represent if the feature is included or not, however, some kind of randomness is involved and it is very hard to select significant features. In GA, classifier accuracies for feature that are already evaluated are not stored for future retrieval and thus lead to multiple evaluations of the same set of features.

K-nearest neighbor (K-nn) is based on the principle that instances within a dataset that have similar properties exist in close proximity to each other. However, the lack of a principled way of choosing K, except through cross validation or similar computationally expensive procedure; larger storage requirement; larger computational time for classification and sensitivity to the choice of similarity function used to compare instances come under the limitations of K-nn.

Support vector machine (SVM) is one of the emerging supervised machine learning techniques based on statistical learning. But, SVM methods are binary, thus, in multi-class problems one must
Exploration and Exploitation in Parallel Problem Solving: Effect of Imitation Strategy and Network Structure
Hua Zhang and Youmin Xi (2012). *Systems Approaches to Knowledge Management, Transfer, and Resource Development* (pp. 201-215).
www.igi-global.com/chapter/exploration-exploitation-parallel-problem-solving/68219?camid=4v1a

Business Intelligence as a Modern IT Supporting Management of Local Government Units in Poland
www.igi-global.com/article/business-intelligence-as-a-modern-it-supporting-management-of-local-government-units-in-poland/163047?camid=4v1a