PARTIAL DISCHARGE FEATURE SELECTION AND EVALUATION USING AN ENHANCED RECURSIVE FEATURE ELIMINATION (RFE) ALGORITHM

Jean **HERNÁNDEZ**, Jesús **VELAZCO**, Universidad de Los Andes, (Venezuela), hmjeanc@ula.ve, velazco@ula.ve
Joshua **PERKEL**, Nigel **HAMPTON**, NEETRAC, (USA),
joshua.perkel@neetrac.gatech.edu, nigel.hampton@neetrac.gatech.edu

ABSTRACT

This paper presents a novel approach for feature selection and evaluation, i.e. a process for reducing and finding an optimal subset of features from an initial set that describes a known dataset. The initial set is used to classify the data into groups, the optimal subset of features disregard unnecessary features that are redundant, which results in better understanding of the classification problem. The approach is divided into two portions, which are an initial selection and a final evaluation. Particularly, the selection portion of the approach is accomplished by performing a cluster variable analysis of the features while the evaluation portion of the process is accomplished by performing an innovative feature evaluation method named recursive feature elimination (RFE) based on support vector machines (SVM). The combined use of the cluster variable analysis and the RFE algorithm is defined as an enhanced-RFE algorithm.

KEYWORDS

Feature selection and evaluation, Support vector machines (SVM), Classification.

INTRODUCTION

Many problems in engineering can be seen as classification problems in which the goal is to use the available information on different items to show how those items relate to each other. The information comes in the form of “features” which are then used as the metrics for making comparisons between the items. Groups of similar items (at least from the perspective of the available features) may then be created. For example, one possible application is classifying different pieces of equipment based on diagnostic testing data into “good reliability” and “poor reliability”. The “good reliability” equipment is left alone while the “poor reliability” equipment is scheduled for maintenance or replacement. On the face of it, this task seems relatively straightforward until one realizes that there is often a large number of features available and that many of them provide no useful information for accomplishing the objective. Therefore, the engineer must:

(1) Decide what features to consider and

(2) Determine which features are the most relevant for the classification problem.

These two processes are often termed feature selection and feature evaluation where the overall goal is to identify a subset of available features that can then be used to perform “successful” classification. This paper presents an approach for completing these processes using diagnostic measurement data.

The approach is divided into two portions: (1) initial feature selection and (2) final evaluation. The selection portion of the approach is accomplished by performing a cluster variable analysis [1] of the features while the evaluation performed using an innovative feature evaluation method named Recursive Feature Elimination (RFE). The RFE method is implemented using Support Vector Machines (SVM) [2]. The use of these two techniques together is the main contribution of this paper and it is defined as the enhanced-RFE algorithm. The paper shows that the algorithm is able to successfully select and evaluate a set of diagnostic features. Specifically, this is illustrated using diagnostic measurements made on medium voltage power cables [3].

FEATURE SELECTION AND EVALUATION

Feature selection and evaluation has been an active research area in pattern recognition, statistics, and data mining applications. The main idea of feature selection and evaluation is to choose a subset of input variables by eliminating features with small or no predictive information. Feature selection and evaluation may improve the understanding of classifier models and often build a model that generalizes better to undetected points. Furthermore, it is often the case that finding the correct subset of features is an important problem in its own right.

Feature selection in supervised learning has been well studied, where the main goal is to find a feature subset that produces higher classification accuracy. Several researches [2, 4-5] have studied feature selection and clustering together with a single goal. For feature selection in unsupervised learning, learning algorithms are designed to find natural grouping of the data in the feature space. Thus, feature selection in unsupervised learning aims to find a good subset of features that forms high quality of clusters for a given number of desired clusters. However, the traditional approaches to feature selection with single evaluation criterion have shown limited capability in terms of knowledge discovery and decision support. This is because decision-makers should take into account multiple-conflicted goals simultaneously. In particular, no single criterion for unsupervised feature selection is best for every application [5] and only the decision maker can determine the relative weights of criteria for the application under consideration.

In general, feature selection methods can be categorized by three types: filters, wrappers and embedded methods [4-5]. Filter methods select subsets of features in terms of criterion functions that are independent of the final classifier used for classification, i.e. feature selection in unsupervised learning. Both embedded and wrapper methods, on the other hand, perform feature selection in the context of learning machines, i.e. feature selection in supervised learning. In embedded methods, feature