A Cosine-Similarity Mutual-Information Approach for Feature Selection on High Dimensional Datasets

Vimal Kumar Dubey, Guru Ghasidas Vishwavidyalaya, Bilaspur, India
Amit Kumar Saxena, Guru Ghasidas Vishwavidyalaya, Bilaspur, India

ABSTRACT

A novel hybrid method based on Cosine Similarity and Mutual Information is presented to find out relevant feature subset. Initially, the supervised Cosine Similarity of each feature is calculated with respect to the class vector and then features are grouped based on the obtained cosine similarity values. From each group the best mutual informative feature is selected. The selected features subset is tested using the three classifiers namely Naïve Bayes (NB), K-Nearest Neighbor and Classification and Regression trees (CART) for getting classification accuracy. The proposed method is applied to various high dimensional datasets. Obtained results showed that the proposed method is capable of eliminating the redundant and irrelevant features.

KEYWORDS

Classification, Cosine Similarity, Feature Selection, Gram-Schmidt Orthogonalisation, Mutual Information

INTRODUCTION

High-dimensional datasets like colon, prostate, and others have the property that numbers of patterns are very less compared to the number of features. Classification (or prediction) of such high-dimensional datasets is very problematic due to a number of features, hence in the recent decades, researchers have focused on feature selection techniques with more intensity. Classification is an indispensable part of data mining (Han & Kamber, 2006), machine learning (Mitchell, 1997) or pattern recognition (Duda et al., 2001) and it is defined as labeling of an unseen pattern based on some information or rules. Classifiers like Support Vector Machines (Cortes & Vapnik, 1995), Naive Bayes (John & Langley, 1995), Artificial Neural Networks (Haykin, 1999) and others take training data and train themselves when an unseen pattern is given to them they labeled those patterns. Since classifiers performance is dependent upon the training data hence elimination of irrelevant, redundant and noisy features is very necessary. Classification accuracy can be increased if non-redundant, relevant and noise free dataset is used for learning. On the contrary, if irrelevant, redundant and noisy features are present in the dataset, it will decrease the classifier performance (accuracy commonly) and often it is termed as Curse of Dimensionality (Jain & Zongker, 1997). Removing irrelevant, redundant and noisy features is termed as Dimensionality Reduction (Liu & Motoda, 1998; Guyon et al., 2006; Saxena et al., 2009). Feature selection (Liu & Motoda, 1998; Saxena et al., 2009) and feature extraction are two well-known methods applied for the dimensionality reduction problem. Processing the existing features of the dataset to obtain new features is termed as feature extraction while the selection of a subset of features from existing set of features without a single extra effort is known as feature selection. In this paper, a novel method is proposed to achieve feature selection.
in databases. The extensive applications of feature selection can be found in medical (Artioli et al., 1995) data mining (Han & Kamber, 2006), classification and other areas (Maimon & Rokach, 2010).

In this proposed method, cosine similarity and mutual information is hybridized to remove the redundant and irrelevant features. Supervised cosine similarity is used as a measure to group the features and then information gain is used to select a best relevant feature from each group. This method provides a way to remove the redundant features.

This paper is organized into following sections. Section 2 introduces related works on feature selection algorithms. Some preliminaries about the terms used in the paper at later stages viz. A measure of similarities and Mutual Information or information gain is given in Section 3. The proposed method is explained via algorithms and model in Section 4. Section 5 lists and briefly explains datasets used in the experiment and experiments. Section 6 and 7 contains results derived from proposed method on listed datasets with a comparison with other methods. Section -8 concludes the paper with the future scope.

RELATED WORKS

Sequential backward selection (SBS) (Marill & Green, 1963) and sequential forward selection (SFS) (Whitney, 1971) feature selection methods had been developed. The nesting effect was the main problem with these feature selection methods. The filter-based feature selection methods and the wrapper-based feature subset selection methods (Bolon-Canedo et al., 2014; Kohavi & John, 1997) are the very efficient method to select relevant feature subset from the feature space. Fast correlation-based feature selection (FCBF) (Yu & Liu, 2003) ReliefF (Kononenko, 1994) CFS (Hall, 1999) and Mutual Information (Peng et al., 2005) are some filter-based feature selection methods. One of the properties of these methods that individual property of features is used to decide whether the feature is relevant or not. In Wrapper (Bolon-Canedo et al., 2014; Kohavi & John, 1997) methods, and all possible feature subset of the original feature set are taken as a possible candidate and the subset which gives best objective value (classification accuracy in the case of classification problem) is taken as the final solution. The above said filter and wrapper techniques can have irrelevance (among the features) and redundancy. (Xie & Wang, 2011; Peng et al., 2010) included an F-score based feature ranking in a sequential forward search method which can eliminate the redundancy among the features. While (Lee et al., 2006) incorporated a Fisher criterion based ranking inside a Binary Search. ReliefF is used by (Zhang et al., 2003) with the genetic algorithm. (Kabir et al., 2012) used the information gain of the features to modify an ant’s position in the Ant Colony Optimization Algorithm. Sequential Random k-Nearest Neighbor (SRKNN) feature selection technique based on wrapper feature selection approach is developed by (Park & Kim, 2015) and applied to high dimensional datasets. In this approach, K bootstrapped samples are generated and evaluated using K-Nearest Neighbour (KNN) (Breiman, 2001) Classifier for the goodness of each sample. It is compared with Backward Elimination Random forest classifier accuracy (Park & Kim, 2015) as the random forest is itself an ensemble of trees. Undoubtedly, there are several relevant papers that use hybrid Feature Selection techniques based on stochastic search methods and on other methods. Saxena et al. proposed an unsupervised feature selection algorithm using Sammon’s stress function and evolutionary method (Saxena et al., 2010). In (Akusok et al., 2016) ELMVIS a model is proposed that used cosine distance in visualization technique for extreme learning machines machine. This method is applied to the face datasets. Cosine similarity is mainly used in the area of document clustering (Al-Anzi & Zeina, 2016). Here we are using cosine similarity to find out the redundancy among the features. The cosine similarity is mainly used in Gram-Schmidt orthogonal feature selection method in which cosine similarity is used in forward features selection and backward elimination approach (Hazewinkel, 2001; Mao, n. d.). A sequential cosine similarity based feature selection method with multiple classifier system is proposed for high-dimensional datasets (Dubey & Saxena, 2015).
An Empirical Investigation of Extensible Information Sharing in Supply Chains: Going Beyond Dyadic
www.igi-global.com/article/an-empirical-investigation-of-extensible-information-sharing-in-supply-chains/119482?camid=4v1a

The Challenge of Relating IS Research to Practice
www.igi-global.com/article/challenge-relating-research-practice/51045?camid=4v1a