Chapter 2
Feature Selection Techniques in High Dimensional Data With Machine Learning and Deep Learning

Bhanu Chander
https://orcid.org/0000-0003-0057-7662
Pondicherry University, India

ABSTRACT

High-dimensional data inspection is one of the major disputes for researchers plus engineers in domains of deep learning (DL), machine learning (ML), as well as data mining. Feature selection (FS) endows with proficient manner to determine these difficulties through eradicating unrelated and outdated data, which be capable of reducing calculation time, progress learns precision, and smooth the progress of an enhanced understanding of the learning representation or information. To eradicate an inappropriate feature, an FS standard was essential, which can determine the significance of every feature in the company of the output class/labels. Filter schemes employ variable status procedure as the standard criterion for variable collection by means of ordering. Ranking schemes utilized since their straightforwardness and high-quality accomplishment are detailed for handy appliances. The goal of this chapter is to produce complete information on FS approaches, its applications, and future research directions.

INTRODUCTION

It is the era of big data, where vast amount of high-dimensional data turn out to be omnipresent in a mixture of fields like online education, social media, bioinformatics, and healthcare. The rapid enlargement of data show disputes for effectual and proficient data organization. It is advantageous to concern data-mining and machine-learning practices to involuntarily determine facts from data of different sorts. But in Data mining and ML fields, while dimensionality of data elevates, the size of data required offering consistent analysis raises exponentially. From the past two decades data demanding appliances

DOI: 10.4018/978-1-7998-6659-6.ch002

demand has increased more and more in terms of capacity, superiority and the extraction of valuable knowledge from such a huge amount of data is not an easy assignment. With this tremendous development, new modern technologies and internet technology applications create a huge amount of data which is unpredicted such as audio, video, text documents, voice conversations, etc, These collected data may restrain high characteristics of measurements that pose a dispute to data scrutiny and result making. With the service of some machine learning (ML) techniques it is possible to reduce, compress the data however still the high dimensional section is momentous issues in mutually supervised and unsupervised ML techniques, now it turns to even more essential with an explosion of available data with the size of data samples along with many relevant features in each sample (Phinyamork et al., 2012; kendall et al., 2015; Barry et al., 2015).

Feature Selection (FS) is the procedure of choosing the most correlated feature points in a data sample, which is essential in ML as well as data mining techniques. Since unimportant or unnecessary features reduce the training speed, interoperability, more importantly, shrinks the generalization performances on the data set. The foremost enthusiasm in dimensionality reduction is to decrease the number of features as low as possible that diminish the training time and increase the taxonomy techniques precision. In detail, feature selection means it is the process of automatically obtaining a subset of features according to employed feature criteria that contribute most to the predicted output which is interested in us. The feature is a variable from the given input data that efficiently express the input at the same time as individual computable belongings of any progression being observed, utilizing those feature sets any ML techniques can perform classification. The focal point of FS is to choose a subset of eliminating noise-related information and produce good prediction results. Feature selection methods able to preprocess learning algorithms, high-quality feature selection consequences which can develop learning precision, condense learning time as well as make things easier to learning results (Bolon et al., 2015; Hira et al., 2015; Pui et al., 2017).

Based on various literature surveys, feature selection techniques foundation on statistics, rough set information theory, and manifold. Numerous techniques are come into action to resolve the problems generated by inappropriate and unneeded variables which are trouble with challenging responsibilities. Depend on the characteristics of feature selection, it applicable in many appliances such as text mining, information retrieval, fault diagnosis, image recognition, bio-metrical data analysis, outlier detection, pattern recognition, data mining, machine learning, and natural language processing, etc. In most of these scenarios feature selection applied at data preprocessing before apply any classification training algorithm. Hence, it also acknowledged as variable selection, variable subset selection, or feature reduction. Standardized data contain thousands of variables where lots of them might be highly associated with additional variables when there two variables are associated with every other merely solitary feature is enough to illustrate the data. Here, the dependent relative variables do not offer any further information regarding the data classes means the entire substance can be acquired from a smaller number of exclusive features that hold the utmost intolerance information on the data classes. Thus removing the reliant variables, the quantity of data can be condensed direct to improvising the categorization performance. In a few situations variables that don't contain a relationship to the classes supply as uncontaminated noise which may initiate bias in the forecaster and diminish the taxonomy presentation. Elimination of irrelevant data needs not to compare with other dimension reduction methods, for the reason that good features independent from the rest of data classes (Jin et al., 2011; Girsh et al., 2017; Han et al., 2011; Song et al., 2007; Starogzyk et al., 2012; Aravjo et al., 2017).