A Survey on Feature Selection Techniques

Jesna Jose[1]

[1]P.G. Scholar, Department of Computer Science and Engg, Sree Buddha College of Engg, Alappuzha

E-mail- jesnaakshaya@gmail.com

Abstract— Feature selection is a term commonly used in data mining to describe the tools and techniques available for reducing inputs to a manageable size for processing and analysis. Feature selection implies not only cardinality reduction, which means imposing an arbitrary or predefined cutoff on the number of attributes that can be considered when building a model, but also the choice of attributes, meaning that either the analyst or the modeling tool actively selects or discards attributes based on their usefulness for analysis. Feature selection is an effective technique for dimension reduction and an essential step in successful data mining applications. It is a research area of great practical significance and has been developed and evolved to answer the challenges due to data of increasingly high dimensionality. The objective of feature selection is three fold. Improving the prediction performance of the predictors, Providing faster and more cost effective prediction and providing a better understanding of the underlying process that generate the data. This paper is actually a survey on various techniques of feature selection and its advantages and disadvantages.

Keywords— Feature selection, Graph based clustering, Redundancy, Relevance, Minimum spanning tree, Symmetric uncertainity, correlation

.

INTRODUCTION

Data mining is a form of knowledge discovery essential for solving problems in a specific domain. As the world grows in complexity, overwhelming us with the data it generates, data mining becomes the only hope for elucidating the patterns that underlie it [1]. The manual process of data analysis becomes tedious as size of data grows and the number of dimensions increases, so the process of data analysis needs to be computerized. Feature selection plays an important role in the data mining process. It is very essential to deal with the excessive number of features, which can become a computational burden on the learning algorithms as well as various feature extraction techniques.. It is also necessary, even when computational resources are not scarce, since it improves the accuracy of the machine learning tasks.This paper made a survey on various existing feature selection techniques.

SURVEY

1. Efficient Feature Selection via Analysis of Relevance and Redundancy

This paper[4] propose a new framework of feature selection which avoids implicitly handling feature redundancy and turns to efficient elimination of redundant features via explicitly handling feature redundancy. Relevance definitions divide features into strongly relevant, weakly relevant, and irrelevant ones; redundancy definition further divides weakly relevant features into redundant and non-redundant ones. The goal of this paper is to efficiently find the optimal subset. We can achieve this goal through a new framework of feature selection (figure 1) composed of two steps: first, relevance analysis determines the subset of relevant features by removing irrelevant ones, and second, redundancy analysis determines and eliminates redundant features from relevant ones and thus produces the final subset. Its advantage over the traditional framework of subset evaluation lies in that by decoupling relevance and redundancy analysis, it circumvents subset search and allows a both efficient and effective way in finding a subset that approximates an optimal subset. The disadvantage of this technique is that it does not process the image data.

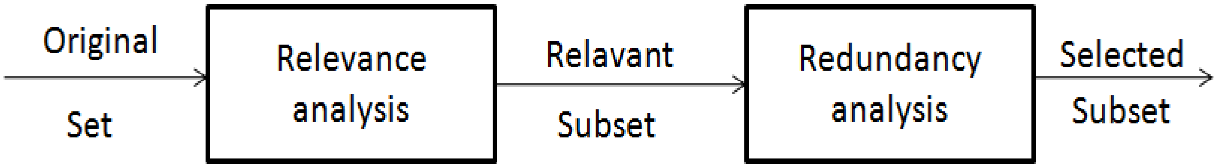

Figure 1: A new framework of feature selection

2. Graph based clustering

The general methodology of graph-based clustering includes the below given five part story[2]:

(1) Hypothesis. The hypothesis can be made so that a graph can be partitioned into densely connected subgraphs that are sparsely connected to each other.

(2) Modeling. It deals with the problem of transforming data into a graph or modeling the real application as a graph by specially designating the meaning of each and every vertex, edge as well as the edge weights.

(3) Measure. A quality measure is an objective function that rates the quality of a clustering. The quality measure will identify the cluster that satisfy all the desirable properties.

(4) Algorithm. An algorithm is to exactly or approximately optimize the quality measure. The algorithm can be either top down or bottom up.

(5) Evaluation. Various metrics can be used to evaluate the performance of clustering by comparing with a "ground truth" clustering.

Graph-based Clustering Methodology

We start with the basic clustering problem. Let X = {x1,…, xNN} be a set of data points, S=(Sij)i,j=1,…,NN be the similarity matrix in which each element indicates the similarity sij ≥ 0 between two data points xi and xj. A nice way to represent the data is to construct a graph on which each vertex represents a data point and the edge weight carries the similarity of two vertices. The clustering problem in graph perspective is then formulated as partitioning the graph into subgraphs such that the edges in the same subgraph have high weights and the edges between different subgraphs have low weights.

A graph can be represented in such a way that A graph is a triple G=(V,E,W) where V = {v1,…, vN} is a set of vertices, E⊆V×V is a set of edges, and W = (Wij)i,j=1,…,N is called adjacency matrix in which each element indicates a non-negative weight ($wwiiii$ ≥ 0) between two vertices vi and vj. The hypothesis behind graph-based clustering can be stated in the following ways[2]. First is the graph consists of dense subgraphs such that a dense subgraph contains more well connected internal edges connecting the vertices in the subgraph than cutting edges connecting the vertices across subgraphs. Second is a random walk that visits a subgraph will likely stay in the subgraph until many of its vertices have been visited (Dongen, 2000). Third hypothesis is among all shortest paths between all pairs of vertices, links between different dense subgraphs are likely to be in many shortest paths (Dongen, 2000)

While considering the modeling step, Luxburg (2006) stated three most common methods to construct a graph: $\varepsilon\varepsilon$ –neighborhood graph, kk-nearest neighbor graph, and fully connected graph. About measuring the quality of a cluster, it is worth noting that quality measure should not be confused with vertex similarity measure where it is used to compute edge weights. The main difference is that cluster quality measure directly identifies a clustering that fulfills a desirable property while evaluation measure rates the quality of a clustering by comparing with a ground-truth clustering.

Graph based clustering algorithms can be divided into two major classes: divisive and agglomerative. In the divisive clustering class, we categorize algorithms into several subclasses like cut-based, spectral clustering, multilevel, random walks, shortest path. Divisive clustering follows top-down style and recursively splits a graph into various subgraphs. The agglomerative clustering works bottom-up and iteratively merges singleton sets of vertices into subgraphs. The divisive and agglomerative algorithms are also called hierarchical since they produce multi-level clusterings, i.e., one clustering follows the other by refining (divisive) or coarsening (agglomerative). Most graph clustering algorithms ever proposed are divisive.

3. Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution

Symmetric uncertainty is in fact the measure of how much a feature is related to another feature. This correlation based filter approach is making use of this symmetric uncertainty method. This involves two aspects: (1) how to decide whether a feature is relevant to the class or not; and (2) how to decide whether such a relevant feature is redundant or not when considering it with other relevant features. The solution to the first question can be using a user- defined threshold SU value, as the method used by many other feature weighting algorithms (e.g., Relief). The answer to the second question is more complicated because it may involve analysis of pairwise correlations between all features (named F-correlation), which results in a time complexity of O(N2) associated with the number of features N for most existing algorithms. To solve this problem, FCBF algorithm is proposed. FCBF means Fast Correlation-Based Filter Solution[3]. This algorithm involves two steps. First step is select relevant features and arrange them in descending order according to the correlation value. Second step is remove redundant features and only keeps predominant ones.

For predominant feature selection another algorithm is there.

a) Take the first element Fp as the predominant feature.

b) Then take the next element Fq.

- if Fp happens to be redundant peer of Fq, remove Fq

c) After one round of filtering based on Fp , take the remaining features next to Fp as the new reference and repeat.

d) The algorithms stops until there is no more feature to be removed.

The disadvantage of this algorithm is that it does not work with high dimensional data.

CONCLUSION

Feature selection is a term commonly used in data mining to describe the tools and techniques available for reducing inputs to a manageable size for processing and analysis. Feature selection implies not only cardinality reduction, which means imposing an arbitrary or predefined cutoff on the number of attributes that can be considered when building a model, but also the choice of attributes, meaning that either the analyst or the modeling tool actively selects or discards attributes based on their usefulness for analysis. Feature selection techniques has wide variety of applications in data mining, digital image processing etc. Various feature selection techniques and its advantages as well as disadvantages are depicted in this paper.

REFERENCES:

[1] I.H. Witten, E. Frank and M.A. Hall, *Data mining practical machine learning tools and techniques*, Morgan Kaufmann publisher, Burlington 2011

[2] Zheng Chen, Heng Ji, *Graph-based Clustering for Computational Linguistics: A Survey* ,Proceedings of the 2010 Workshop on Graph-based Methods for Natural Language Processing, ACL 2010, pages 1–9, Uppsala, Sweden, 16 July 2010. c 2010 Association for Computational Linguistics

[3] Lei Yu, Huan Liu, *Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution*, Department of Computer Science & Engineering, Arizona State University, Tempe, AZ 85287-5406, USA

[4]Lei Yu, Huan Liu, *Efficient Feature Selection via Analysis of Relevance and Redundancy*, Journal of Machine Learning Research 5 (2004) 1205–1224