Feature Selection: Concept, Principle and Technical Issues

Wenxue Huang[1]

Department of Mathematics

Shantou University

E: whuang123@yahoo.com

Extended Abstract: Suppose we are given a high dimensional data set consisting of tens, hundreds, even thousands of independent variables and a dependent variable and with large sample size. We are also given an explicit or implicit quantitative project objective to optimize based on the given information.

There can be (actually, is most likely) a lot of noisy or redundant information in the data set for our project (prediction, controlling, clustering or just finding causal factors), which makes the optimization challenging.

Data can be static or dynamic, even longitudinal; data can be of mixed type: variables are of types of continuously or discretely interval-scaled or ratio-scaled, nominally and ordinally categorical; even a variable itself can be of nature of different types.

To realize the optimization based on the data, we need to appropriately reduce the dimensionality and find out a small number (usually less than 10) of independent variables as a base (like a base for a vector space). The process of finding is referred to as feature selection or dimension reduction.

Feature selection is an important but also quite popular topic in data mining and has a long history. When data is of a single type, objective is of typical nature, and when dimension is not quite high

[1] Tel. :+86-754-8290-3745

(say, < 1000), in most cases, there are corresponding feature selection methods or models, of which some are quite robust while others not so.

A nature question arises. Is there any universal feature selection approach which may quite robustly or reliably handle the above mentioned complicated situations? This is the concern of this presentation.

I'm going to discuss the basic concept, principle and technical issues in general feature selection approaches, and provide a specific approach which is based on categorical data but can apply to or be easily adapted to general situations. If the dependent variable itself is categorical, or the dependent variable is continuous and (often) can be (conventionally or technically) rationally discretized, this approach I believe is still quite robust. Otherwise more care is needed.

When dependent variables are also of high dimension, how to do feature selection? I'm going to briefly discuss this issue as well.

Keywords: data types, feature selection, project objective oriented.