Feature Selection with Mutual Information for Regression Problems.

Muhammad Aliyu Sulaiman[1], Jane Labadin[2]
Faculty of Computer Science and Information Technology, Universiti Malaysia Sarawak,
94300 Kota Samarahan, Sarawak, Malaysia
[1]muhalisu@gmail.com , [2]ljane@pps.unimas.my

***Abstract*— selecting relevant features for machine learning modeling improves the performance of the learning methods. Mutual information (MI) is known to be used as relevant criterion for selecting feature subsets from input dataset with a nonlinear relationship to the predicting attribute. However, mutual information estimator suffers the following limitation; it depends on smoothing parameters, the feature selection greedy methods lack theoretically justified stopping criteria and in theory it can be used for both classification and regression problems, however in practice more often it formulation is limited to classification problems. This paper investigates a proposed improvement on the three limitations of the Mutual Information estimator (as mentioned above), through the use of resampling techniques and formulation of mutual information based on differential entropic for regression problems.**

Keywords—Mutual Information; Feature Selection; Regression Problems

I. INTRODUCTION

More often, many applications generate datasets with large number of attributes/variables. These datasets are not necessarily meant to be used for machine learning predictions. As a result of that, some of the variables may be irrelevant to the predicting attribute(s). And their presence in the set may affect the predicting capability of a machine learning model.

Feature selection aims at reducing the dimension of a dataset by selecting variables that are relevant to the predicting attribute(s). And this helps to improve the predicting capabilities of the machine models in the following ways; (1) Selected feature subset helps in building concise model which often avoid over-fitting and generalized better. (2) Feature subset selection can improve accuracy of prediction because of reduction in estimation error. (3) Building good predictor model often requires reduction in feature subset. (4) Feature subset selection reduces the burden on data collection and as well reduces computational complexity.

This paper investigates on how the feature selection based on mutual information is extended to regression problems. The rest of the paper is organized as follows: Section II is a review of information theory concepts as it relates to mutual information and how mutual information can be used as a relevant criterion for feature selection. Section III provides details on how mutual information criterion is estimated for regression problem and improvement in feature selection algorithm. Section IV contains experimental studies, including experimental results and discussion. Section V presents conclusion and future work.

II. REVIEW

A. Entropy and Mutual Information (MI)

Entropy of a random variable X is an information theory concept that measures the uncertainty associated with X. Whereas Mutual information (*MI*) which is another information theory that quantitatively measures the amount of dependent information two random variables have about each other. Unlike correlation coefficient that measures linear dependence only, mutual information measures both linear and nonlinear dependence between variables, a property that made it a popular choice for feature selection [1, 2, 3, 4 and 5]. Let us consider a pair of continuous random variables X and Y, the joint probability density function of X and Y is expressed as:

$$P_{X,Y}(x, y) = P_Y(y|x)P_x(x) \quad (1)$$

In a similar way the joint differential entropy of X and Y is expressed as:

$$h(X, Y) = h(X) + h(Y|X) \quad (2)$$

Where $h(Y|X)$ is known as the conditional differential entropy of Y given X. In word the equation 2 is expressed as the uncertainty about X and Y is equal to the uncertainty about X plus the uncertainty about Y given X. This can equally be said in the other way round as the uncertainty about X and Y is equal to the uncertainty about Y plus the uncertainty about X given Y as in equation 3:

$$h(X, Y) = h(Y) + h(X|Y) \quad (3)$$

Meanwhile, entropy of a random variable X is expressed as

$$h(X) = -\int f_X(x) \log f_X(x) d_x \quad (4)$$