Computational Statistics and Data Analysis 72 (2014) 105–127

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Time-efficient estimation of conditional mutual information for variable selection in classification

Diman Todorov*, Rossi Setchi

Knowledge Engineering Group, Institute of Mechanical and Manufacturing Engineering, Cardiff University, UK

ARTICLE INFO

Article history:
Received 29 June 2012
Received in revised form 28 October 2013
Accepted 31 October 2013
Available online 8 November 2013

Keywords:
Variable selection
Conditional mutual information
Discrete data
Parallel algorithm

ABSTRACT

An algorithm is proposed for calculating correlation measures based on entropy. The proposed algorithm allows exhaustive exploration of variable subsets on real data. Its time efficiency is demonstrated by comparison against three other variable selection methods based on entropy using 8 data sets from various domains as well as simulated data. The method is applicable to discrete data with a limited number of values making it suitable for medical diagnostic support, DNA sequence analysis, psychometry and other domains.

1. Introduction

Variable selection is a family of dimensionality reduction methods used to identify predictors with the best prediction performance. Unlike projection based methods for dimensionality reduction, such as principal components analysis, the result of variable selection is easily interpretable in the context of the application domain.

Many variable selection methods use an estimate of a correlation coefficient $\rho(Y, X)$ between a response variable Y and a candidate predictor variable X where the coefficient is calculated by assuming a distribution in the data—usually a normal distribution. The assumption of normality can be alleviated for discrete data by assuming that the data comes from an underlying discretised normal distribution (Kolenikov and Angeles, 2004). However this assumption cannot always be made and is difficult to test for when the data is discrete.

Other methods use an estimate of the difference of two distributions p and q without making an assumption about the distributions. The difficulty with the latter approach is that while it has the advantage that it is more flexible with regard to the shape of the distributions, it requires more data for accurate estimation of the distributions.

The general form of a method of this kind is a distance function D of the form:

$$D(p \parallel q). \tag{1}$$

In the context of this work, p is the distribution $p(Y, X)$ and $q = p(Y)p(X)$. To adapt this form for minimising the correlation of X_1 with another predictor X_2 while maximising the correlation with Y, the form can be modified so that $p = p(Y|X_2, X_1|X_2)$ and $q = p(Y|X_2)p(X_1|X_2)$.

This modified method however still only accounts for redundancy between pairs of predictors. The naive approach to estimate all redundancy of X_1 as a predictor of Y is to estimate D while replacing X_2 for all subsets of predictors which do not include X_1. While this would account for all redundancy, the benefit comes at a cost: the computation is subject to

* Corresponding author. Tel.: +43 6763354582.
E-mail address: todorovd@cardiff.ac.uk (D. Todorov).

0167-9473/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.csda.2013.10.026

combinatorial explosion in the number of subsets of predictors; the computation requires the estimation of multivariate probabilities from data.

Several functions have been proposed as an estimate of the distance between a response and a predictor for feature selection (Guyon et al., 2006, p. 96 and following). The research presented in the following is focused on "*distance*" measures based on entropy. The reason for investigating information theory – and conditional mutual information in particular – is that it provides a comprehensive framework for estimating differences between variables while taking conditional probability into account (Cover and Thomas, 1991).

In literature the estimation of conditional mutual information is often dismissed as computationally infeasible for practical applications (Liu et al., 2009). Computationally efficient methods for estimating conditional mutual information have been proposed (Torkkola, 2003; Jung et al., 2011; Tsimpiris et al., 2012), however they often are applicable to discrete data only with limiting assumptions. Jung et al. (2011) propose a method which incorporates conditional mutual information. The innovation of the method is in the estimation of probabilities using a Gaussian mixture model. While GMMs can be identified for ordinal data (Iannario, 2010) the method is rather complex. Tsimpiris et al. (2012) acknowledge that their method is not suitable for discrete data. The authors propose extending discrete features with white noise before applying their method.

An example of an algorithm which limits the application domain to reduce computational complexity has been proposed by Fleuret (2004). Their algorithm is limited to binary variables and two classes. Although this limitation may seem overly restrictive, binary data can be found in many domains of research (de Leeuw, 2006).

In this paper an algorithm is proposed which calculates the classical estimator for conditional mutual information by utilising the massively parallel computation hardware on modern graphics cards. The algorithm can process discrete data in multiple classes – as opposed to only binary data in two classes – and allows an exhaustive search in the subset space of candidates. The main difficulty in this approach becomes the availability of sufficient amounts of data needed to estimate high dimensional probability functions.

2. The information theoretic approach to variable selection

In this section the foundations of information theory are introduced and state-of-the-art methods for information theoretic variable selection are discussed. Finally, the theoretical foundations of the method proposed in this paper are presented.

2.1. Information theory and feature selection

Information theory is concerned with measuring the amount of information transferred between an information source and information sink. It has applications in many fields not related to communication, for example physics and cryptography. The measure used in information theory is entropy, defined as:

$$H(X) = -\sum_{u} p(X_{X=u}) \log(p(X_{X=u})), \tag{2}$$

where X is a random variable, u is a value which X can take and $p(X_{X=u})$ is the probability that X will take the value u. When the logarithm is to the base of 2, entropy is measured in bits. The value of entropy is always greater or equal to zero and its upper limit is bounded by the logarithm of the number of values u which X can take. When the value is high, a lot of information is obtained by observing X. If on the other hand the value is zero, when there is only one value u with $p(X_{X=u}) = 1$, no information is obtained by observing X.

For developing a variable selection method based on entropy it is of interest to know how much information a predictor variable X contributes to an outcome variable Y. This magnitude can be modelled with conditional entropy, which is defined as:

$$H(Y|X) = -\sum_{u,v} p(Y_{Y=u}, X_{X=v}) \log(p(Y_{Y=u}|X_{X=v})), \tag{3}$$

where u are values which Y can take, v are values which X can take and $p(Y_{Y=u}|X_{X=v})$ is the probability that Y will take the value u when X is known to have taken the value v.

Using conditional entropy, a measure can be defined, which represents the distance between the probability distributions of X and Y. This distance measure is called mutual information and is defined as:

$$I(Y; X) = H(Y) - H(Y|X). \tag{4}$$

Mutual information, $I(Y; X)$, is interpreted as the reduction of uncertainty of Y due to the knowledge of X. It is always greater than zero and its upper bound is $\min(H(Y), H(X))$. The value of $I(Y; X)$ is zero when X carries no information about Y. Its value is $H(Y)$ when X perfectly describes Y.