Variable Importance Evaluation for Machine Learning Tasks

Martti Juhola
University of Tampere, Finland

Tapio Grönfors
University of Eastern Finland, Finland

INTRODUCTION

Computational methods are presented to examine the importance of variables in data sets for variable evaluation and weighting and other purposes. Frequently, it is useful to explore for the machine learning tasks whether a data set includes such information that is important for the separation of classes in data. Such an evaluation can be performed for an entire data set or separately classes or variables. We concisely describe basic techniques and principles of more sophisticated but also complicated methods. We describe a recently introduced method called Scatter based on traversing through a data set as near neighbour cases and counting class changes. The fewer the changes, the more compact the classes are in a variable space so that they are possible to separate with high accuracy. We present an example how to use the Scatter method and conclude discussing its usefulness for the evaluation of variables for a data set.

BACKGROUND

Variables or features or attributes of a data set represent the properties of a data case or instance in the context of a phenomenon or an object to be studied. Such an object can vary from concrete visual patterns in digital images to abstract ones as diseases that appear in individual subjects. Variable values are measured in several ways depending on object types as patients, measured in physiological or other laboratory tests or collected with inquiries given to them. Sometimes data values are computed directly from a data source like occurrences of certain words from among an electronic document collection.

Variable types can vary how they are formed. For instance, in images features can be formed on the basis of geometric and statistical approaches. An object in an image is mapped with the length of its periphery, shape as compared to a circle, ellipse or square, its diameter and many other measures originated from its size and shape. Statistical features are associated with the distributions of intensities of grey levels or colours encoded numerically. For example, mean, variance, skewness and kurtosis being statistical moments can be calculated to characterize an object. Further, such features can also be measured that connect to other factors, say, the location of an object in an image.

Variable types are also seen according to how they are represented in computation. Measuring devices typically record signals of continuous phenomena, e.g., temperature, but output discrete values. For instance, an analog-digital converter may give signal data (perhaps voltage first amplified) values between -10 V and +10 V. These values are not continuous, but discrete, because they are digitized, e.g., according to 16 bit words into an interval such as $[0,2^{16}-1]$. This may be calibrated in a way or another to correspond to some property, for example, a subject's weight in kilograms. In the statistical sense variable types are nominal, ordinal, interval or scales ratio (absolute). All of these types may appear within a medical data set. For instance, the colour of eyes is nominal, and the grade of pain is ordinal, say 'no pain', 'slight' or 'severe'. Usually these are encoded with non-negative

DOI: 10.4018/978-1-4666-5888-2.ch029

integers. However, we have to know which statistics are possible to compute for them. For nominal variables we can compute modes. We cannot compute means, standard deviations and several other quantities for them, but only for interval and scales ratio. The difference of the two latter is that interval type has no fixed zero or measuring unit. For example, temperature can be measured in different ways.

Frequently, variables are not of the same importance for computation such as classification between different possible diseases of patients in a given medical specialty. It is perhaps important weight variables or discard less important. A great number of variables are nowadays common for several applications. For example, document classification or text categorization is an area where there may be thousands of variables, representing relative frequencies of different relevant words present in documents. Therefore, we have to somehow reduce huge numbers of variables to enable computation in sensible running times and, in general, to leave out unnecessary variables the usefulness of which is negligible for a current computational task.

To preprocess data such actions as the imputation (substitution) of missing values with their estimates or normalization of different variable scales are frequently required. We assume that these, if necessary, are performed before evaluating the importance of variables. Nevertheless, abundant missing values for certain variables are a natural cause to reject these variables to avoid unreliable decisions with a data set, but this situation is encountered at an early preprocessing stage before using selection or evaluation algorithms of variables.

A good set of variables can be selected either independent of a machine learning algorithm as a result of preprocessing based on the data itself or with a machine learning algorithm, when typically the subset of variables which performs best is selected.

In the following we briefly present some methods for variable selection and the Scatter method for the evaluation of variable importance. As an example, we also present how this method was applied to a medical data set for the purpose of the classification of cases (patients) into a few different disease classes.

METHODS FOR VARIABLE SELECTION AND IMPORTANCE EVALUATION

A simple statistical way to explore variables is to compute correlations between variables. This can be done provided that variables are at least ordinal. Correlation coefficients reveal mere linear relations. There can be non-linear dependencies of higher degrees, virtually unlimited of numbers of degrees (Pyle, 1999). Consequently, correlations or corresponding simple statistical means are fairly seldom sufficient to operate data sets with a great amount of variables. If there are only two classes for a classification task, basic statistical methods can be used more extensively. For instance, t test can then be applicable provided that its conditions are satisfied. A straightforward technique is mutual information based on probabilities. Its basic form is applied to nominal variables, but it is possible to extend for other types (Guyon & Elisseeff, 2003). For nominal values probabilities are estimated with frequencies, but for other variable types their calculation is more complex. Discretizing variable values can be used or, for continuous variables, their densities can be approximated.

There are a great amount of statistical methods to predict by using, e.g., multiple regression models (George & McCulloch, 1993). Such statistical techniques frequently apply continuous (real or scales ratio) input variables and also give a real value output. However, this choice is unsuitable for classification, where c classes are encoded with nominal values $\{1,2,\ldots,c\}$ and these values are only used to act as class labels. Statistical techniques may also be unsuitable for mixed-type variables, where a data set can contain utmost types, from the binary and nominal to continuous.

Importance or relevance of variables can be defined in different ways, e.g., a variable is relevant to a class provided that there is a pair of cases so that these two differ exclusively in the values of this variable and the classes of the two cases (Bell & Wang, 2000; Wolf & Shashua, 2005). We follow a lighter concept by describing quantitatively that a variable is more or less important for classification in a data set given.

Data Recognition for Multi-Source Heterogeneous Experimental Detection in Cloud Edge Collaboratives
Yang Yubo, Meng Jing, Duan Xiaomeng, Bai Jingfenand Jin Yang (2023). *International Journal of Information Technologies and Systems Approach (pp. 1-19).*
www.irma-international.org/article/data-recognition-for-multi-source-heterogeneous-experimental-detection-in-cloud-edge-collaboratives/330986