New Entropy Based Distance for Training Set Selection in Debt Portfolio Valuation

Tomasz Kajdanowicz, Wrocław University of Technology, Poland
Slawomir Plamowski, Wrocław University of Technology, Poland
Przemysław Kazienko, Wrocław University of Technology, Poland

ABSTRACT
Choosing a proper training set for machine learning tasks is of great importance in complex domain problems. In the paper a new distance measure for training set selection is presented and thoroughly discussed. The distance between two datasets is computed using variance of entropy in groups obtained after clustering. The approach is validated using real domain datasets from debt portfolio valuation process. Eventually, prediction performance is examined.

Keywords: Dataset Selection, Debt Valuation, Distance Measures, Intelligent Systems, Prediction Methods, Supervised Learning

INTRODUCTION
Supervised learning’s task is based on assumption that one is able to provide proper training data which will be used for further generalization and inference process. The quality of data directly affects performance of prediction and classification algorithms. Training data consist of a set of training examples composed of a pair of input features $X$ and a desired output value $Y$.

The main aim is to analyze training data and produce an inference function $\Phi$ that maps input to output, $\Phi: X \rightarrow Y$. The output can be twofold: in case of discrete one, the function $\Phi$ is called a classifier. Otherwise, in case of continuous output, it is called a regression. If function $\Phi$ maps to interrelated set of more than one values it is structured prediction or structured output learning algorithm. All in all, the inferred function $\Phi$ should be able to predict the correct output value for any valid input object. This requires learning algorithm to be able to generalize based on the training data. Consequently the quality of data is of great importance. Therefore, the selection of training set for learning algorithm should be performed carefully in order to select optimal dataset. The most straightforward and clear situation arises when learning concerns data from particular domain and describes always the
same stationary object. In such case, properties of data and statistical dependencies between examples remain unchanged and training may be performed using the same source of training and testing data. Such data, as long as being of appropriate size, may deliver satisfactory generalization abilities. In order to generalize from data describing non-stationary objects, learning algorithms are expected to model concept drift (Kurlej & Woźniak, 2011) phenomenon identified by changes in data probability distributions. As concept drift may be caused by changes of prior, conditional or posterior probabilities of data, appropriate methods must be incorporated to address the problem.

Another situation occurs when generalization needs to be performed for objects for which training data is not available or hardly accessible. In such case, learning is performed using data describing other similar objects. An example of such a situation are across-network classification where learning performed on one network adjust models used in generalization on another network (Lu & Getoor, 2003) or debt portfolio value prediction where value of appraisal of particular portfolio is done using other similar portfolios (Kajdanowicz & Kazienko, 2009).

The paper considers the latter problem of training set selection in the prediction task for prediction of future debt recovery. Intuitively, the greater similarity/smaller distance between objects used in learning and those the inference is applied to, the better performance of inference methods. Similarity/distance identification between training and testing objects can be reduced to similarity/distance measurement between datasets describing their input features, namely similarity/distance between \(X_{\text{train}}\) and \(X_{\text{test}}\). Aforementioned similarity and distance can be invoked interchangeably as similarity can be measured by distance, i.e., two objects are similar if the distance is close to zero. Growing distance results in higher dissimilarity.

In general, distance is defined as a quantitative degree of how far apart two objects are (Cha, 2007). The choice of distance measure depends on the representation of objects and type of measurement. In supervised learning tasks datasets are usually represented by matrices in which columns denote attributes and rows object instances. A single cell of such matrix contains a value of particular attribute for a given instance. Hence, the problem of training set selection based on measuring the distance between two datasets \(X_{\text{train}}\) and \(X_{\text{test}}\) is actually a matrix distance based selection.

Altogether, training dataset can be obtained by performing one of the following scenarios (Figure 1):

- Example selection from available historical datasets based on the distance between particular testing and training examples (e.g., Cano, Herrera, & Lozano, 2003; Son & Kim, 2006).
- Training set selection from a set of available historical datasets based on the distance between particular testing and training sets.

This paper concerns the latter approach and introduces a method for training dataset selection. The method is designed to improve inference results using entropy-based distance between probability density functions of training and testing datasets.

The rest of the paper is organized as follows. In section Related Work various approaches and distance measures that may be utilized to training set selection are enumerated. In order to provide a better perspective on the problem, section Debt Portfolio Value Prediction presents a real-world training set selection problem in debt portfolio value prediction. In section Training Set Selection Using Entropy Based Distance a new approach to measure distance for comparison and selection of training datasets is described. Evaluation of the impact on prediction accuracy using proposed method is discussed in section Experiments and Results. Finally, section Conclusion summarizes the work.