Training Set Construction Methods

Tomas Borovicka*
Department of Software Engineering
Faculty of Information Technology
Czech Technical University in Prague
Thakurova 9, 160 00 Praha 6, Czech Republic
borovto1@fit.cvut.cz

Abstract

In order to build a classification or regression model, learning algorithms use datasets to set up its parameters and estimate model performance. Training set construction is a part of data preparation. This important phase is often underestimated in data mining process. However, choose the appropriate preprocessing algorithms is often as important as choose the suitable learning algorithm. Goal of training set construction algorithms is to build representative datasets by discarding useless instances and enforcing important instances. Good quality training set is a good premise to build a well learned and reliable model. Lot of literature have been published about comparison of learning algorithms and regression or classification models, but good review and comparison of training set construction methods have not yet been given. This work is focused on how to select data samples from an original set and place them into the training and testing sets. In the first part is an overview of existing approaches and new possible approaches are discussed. The second part is focused on experimental comparison of these methods.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications–*Data mining*; I.2.m [**Artificial intelligence**]: Miscellaneous; H.3.m [**Information Search and Retrieval**]: Miscellaneous

Keywords

training set construction, classification, regression, representative set, data splitting, instance selection, class balancing, machine learning

*Master study programme in field Software Engineering. Supervisor: Dr. Pavel Kordik, Department of Software Engineering, Faculty of Information Technology, Czech Technical University in Prague.

Borovicka, T. Training Set Construction Methods. Information Sciences and Technologies Bulletin of the ACM Slovakia, Special Section on the ACM Student Project of the Year 2012 Competition, Vol. 4, No. 4 (2012) 43-44

1. Introduction

A training set is a special set of labeled data providing known information that is used in the supervised learning to build a classification or regression model. Each training instance consists of vector of n input attributes $\mathbf{x} = (x_1, \ldots, x_n)$ (features) and an appropriate output value y (response variable). The role of supervised learning algorithms is to produce a function $f(x)$, based on given training set $R = \{\mathbf{x}_i, y_i\}_1^N$, that makes a prediction y' for future data where only values of $\mathbf{x}$ are known. It means that deduced classification or regression function should predict the most likely output value for any input vector. The goal of the training phase is to estimate parameters of a model to minimize inaccuracy between predicted y' and real value y. Capability of the model to predict output value can be evaluated by various measures. General it is called the prediction performance.

A training set R should be a representative set of a population, when we want to built a good quality model. Population is a set of all existing feature vectors (features) and representative means that satisfies these main characteristics [4]:

1. It is significantly smaller in size compared to the population.
2. It captures the most of information from the population compared to any subset of the same size.
3. It has low redundancy among the representatives it contains.

Usually sample set S, subset of a population collected during some sampling process, is available. In ideal case is S representative, but in practise it is unusual. When this dataset is not ideal we can perform some task to make it more representative.

1.1 Data Splitting

In order to assess a model it should be evaluated using appropriate measures. The goal is to estimate future predictive performance on unseen data. For reliable prediction, to avoid over-fitting or to reveal bias, model should be evaluated on independently collected sample (different, independent and identically distributed). Unfortunately, common situation is that we have not more independent datasets and we can not easily collect new ones. In this case we can split a dataset into more disjunctive subsets to simulate the effect of having more independent datasets.

Sets used for an evaluation of a model are the validation set V (usually used for a model selection) and the testing set T (used for model assessment). The question is how to split a dataset into more subsets for learning and evaluation and keep the best possible quality of a model?

1.2 Instance Selection

The size and quality of datasets are different from case to case. Some are small, some large and contain varying amounts of noise and redundancy. Learning algorithms usually use all instances of given training set R during the learning phase even though lot of them are useless and they can not increase predictive performance of a model. There are several reasons to remove them. Besides, they do not increase performance of a model, they can even degrade it. Discarding noisy and redundant instances usually leads to increase in quality of a model. Moreover, the decreasing amount of instances reduces the computation, that can be very complex with huge datasets. The process of discarding useless and enforcing important instances is in literature usually called instance selection.

1.3 Class Balancing

Many datasets from particular domains (very often medical data) are characterized with a few amount of instances in one class, which is usually our point of interest (positive instances), and large amount of instances in other class (negative instances). This datasets are called imbalanced. Most of learning algorithm expect that training sets have same proportion of classes (they are well balanced) and on imbalanced datasets have a poor results. Methods deal with this problem are in literature called class balancing methods.

2. Review

Several data splitting, instance selection and class balancing methods published in literature have been reviewed as well as methods for their evaluation and comparison. Also new approaches have been discussed. For detailed review see [2].

3. Experiments

Experiments are divided into three parts where each part corresponds to one particular task related to training set construction, described above. Experiments are performed in the following way. For each experiment are chosen appropriate evaluation measures. Experiments are designed to assess the behaviour of the algorithms on different datasets and for various classification models. Results are compared in relation with a particular dataset and concrete classification model. All dataset used for benchmarking are from the *UCI Machine Learning Repository* [3]. Datasets were chosen by their characteristics to be appropriate for our experimental purposes. Each dataset have its main characteristics which are crucial for assessment of the results of each experiment.

For data splitting methods no method outperforms other methods on all or most of the datasets, as expected. But interesting findings has been observed. CADEX has had significantly better results than other methods on imbalanced datasets, it shows that more sophisticated methods for data splitting could improve accuracy of a model. it is not surprising that random and stratified sampling have had the most stable results over all datasets, but it is surprising that this naive methods can outperform cross-validation and bootstrap in most cases. The bootstrap has had similar results as the cross-validation but usually with a smaller variance and with a small pessimistic bias. Advantage of the cross-validation and bootstrap is that they use all available data for learning, this gives better results on small datasets.

It has been observed that instance selection algorithms can significantly reduce datasets and keep still good quality of a model. The Decremental Reduction Optimization Procedure (DROP) outperforms other methods on all datasets and in both used measures – model performance on produced subset and the compression rate, which is defined as $compression\ rate = \frac{|R^*|}{|R|}$, where $|R|$ is the size of an original set and $|R^*|$ is the size of a set of selected instances. DROP has reduced amount of instances to less than one fifth with a small decrease in the classification performance on most of datasets. ENN on most of the datasets increased the classification performance of models, it confirms that ENN works well as the noise filter. Also obvious relation between amount of instances and learning time of some models has been showed. The amount of instances in the training set affects the learning and response time that can instance selection methods rapidly reduce.

Class balancing methods increase the performance of classification models on imbalanced datasets. These methods significantly increase sensitivity of a model on the minority class. But unfortunately with increasing sensitivity is usually decreased precision of positive class classification. It means that is a lot of false positive classifications and overall performance can even decrease. As previous methods no class balancing method has had significantly better results than others over all datasets.

Detailed results of all experiments can be found in [1].

4. Conclusions

Several methods for data splitting, instance selection and class balancing published in literature were reviewed. For each group of methods has been created methodology for its comparison using appropriate measures. According to this methodologies have been performed experiments. Described methods can significantly increase classification performance of learned models. All compared methods have had different results on various datasets. This indicates that methods are strongly domain dependent. Moreover, results of methods differ in dependence of used classifier. Comprehensive review of training set construction methods with experimental results and statistically significant findings, that can help researchers to make decision which methods use in their case has been given.

References

[1] T. Borovicka. Training Set Construction Methods. Master's thesis, Czech Technical University in Prague, 2012.

[2] T. Borovicka, M. Jirina Jr., P. Kordik, and M. Jirina. Selecting representative data sets. In A. Karahoca, editor, *Advances in Data Mining Knowledge Discovery and Applications*. Intech, 2012.

[3] A. Frank and A. Asuncion. Uci machine learning repository, 2010.

[4] F. Pan, W. Wang, A. Tung, and J. Yang. Finding representative set from massive data. In *Data Mining, Fifth IEEE International Conference on*, pages 8–pp. IEEE, 2005.