Training Set Construction Methods

Tomas Borovicka supervised by Pavel Kordik

Faculty of Information Technology, Czech Technical University in Prague

Motivation

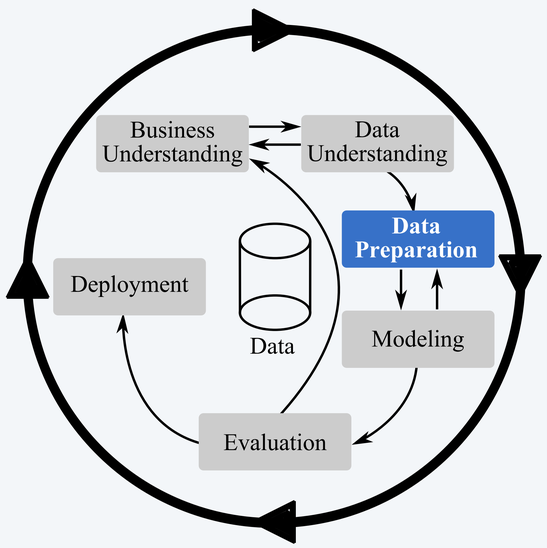

Figure 1: Life cycle by CRISP-DM

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** or even omitted in data mining process. However, choose the appropriate preprocessing algorithm is often more important than chosen learning algorithm or used model.

Goal of these 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 thesis is focused on **how to select data samples from an original set and place them into the training and testing sets**. In the first chapter is an **overview of existing approaches** and new possible approaches are discussed. The second chapter is focused on **experimental comparison** of these methods.

Review and Experiments

Training set construction methods can be divided be their purpose into the three main groups: **data splitting**, **instance selection** and **class balancing**. Each group deals with different problem (Figure 2).

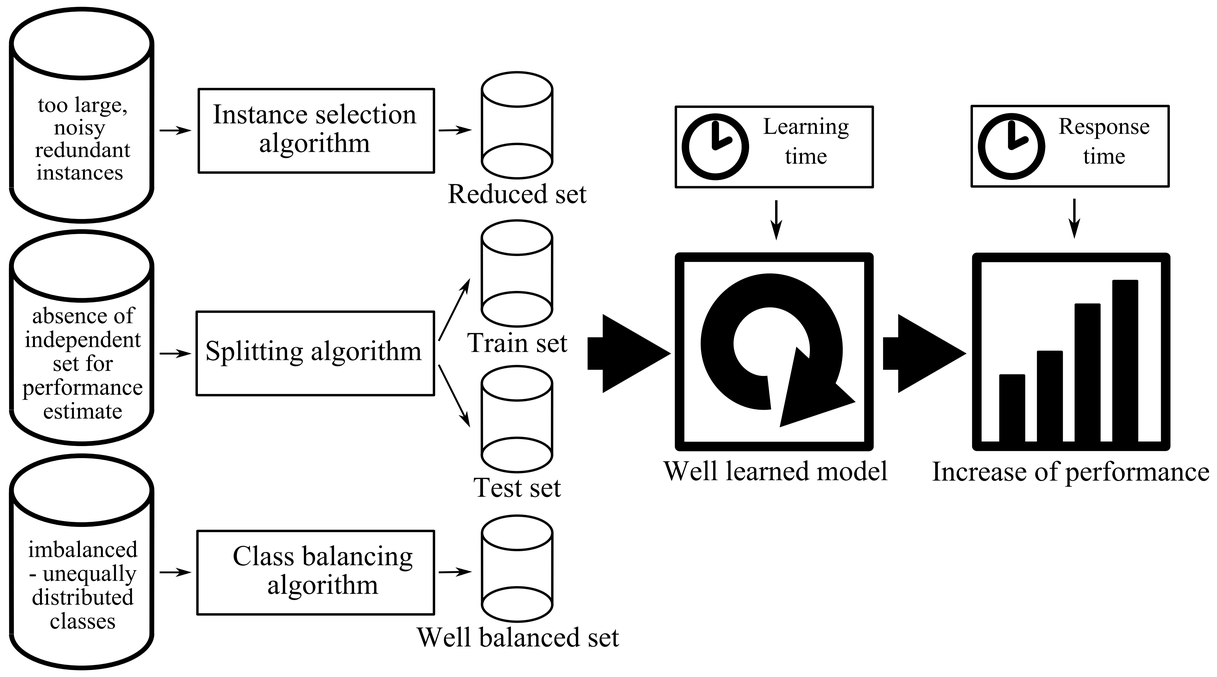

Figure 2: Training Set Construction Methods

Reviewed methods

Several data splitting, instance selection and class balancing methods published in literature have been reviewed and new approaches have been discussed.

- Holdout method
- Repeated holdout
- k-fold cross-validation
- Leave-one-out CV
- .632 bootstrap
- Kennard-Stone
- DUPLEX
- NN splitting
- Closests pairs splitting
- RNN, SNN
- CNN, GCNN
- ENN, RENN, Multiedit
- EECCB
- IB1-3, DROP1-5, ICF
- POP, POC-NN
- Maxdiff kd trees
- NSC, GCM
- NSB, OSC, WP, PSR
- Random under/over-sampling
- NCL, OSS, OSPC
- Tomek Links
- SMOTE, Borderline SMOTE
- Budget-sensitive prog. sampling
- One class learning
- Cost sensitive learning
- AdaBoost, MetaCost, AdaCost
- SMOTEBoost, RareBoost

Goal of Experiments

- Test methods on **various benchmarking datasets**.
- Evaluate classification performance obtained by **different classifiers**.
- **Compare methods** quality using appropriate measures.

Results and Conclusion

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.
- Most commonly used data splitting algorithms, cross-validation and bootstrap, have had poor results with high variance.
- Instance selection algorithms can **significantly reduce training set** and still reach **high performance** on unseen data.
- Moreover, they can significantly **speed up learning phase and response** of built models.
- Class balancing methods increase sensitivity of a model to a minority class, but usually significantly decrease precision.
- All compared methods have had different results on various datasets. This indicates that methods are **strongly domain dependent**. Moreover, results of methods are **dependent on specific classifiers**.
- It is necessary to carefully choose the appropriate method for particular domain and learning algorithm.

Contribution

- Comprehensive review of training set construction methods with experimental results.
- Statistically significant findings, that can help researchers to make decision which methods use in their case.
- Review has been accepted as a chapter in book Data Mining / Book 2, InTech.