Prototype Applications of Instance-Based Reasoning

Jiří Kléma

Thesis abstract

The presented thesis focuses on instance-based learning (IBL) methods. The groundwork of instance-based learning was put in 1950's. Recently, the original idea of the nearest neighbor algorithm has attracted attention of machine learning community. It resulted in many investigations and improvements in the given field. This development was followed by many successful applications in the past years. At the same time, it has been shown that some of the practical applications can be solved by a direct utilization of the given methodology, others call for its further development. The given thesis selects applications belonging to the later group and addresses issues of the problem oriented IBL development mainly.

The thesis is concerned with two particular prototype practical applications. The application character of the thesis qualifies that it does not suggest any single general algorithm but it rather focuses on process of the final task dependent solution design. At the same time, it tries to show that the development of such solutions can bring algorithm modifications and improvements that can be generalized to group of relevant domains. The first presented domain regards a task of gas prediction consumption, the second one deals with early detection of pump faults. These domains represent diverse learning tasks with different singularities, different targets, different feature definition and characteristics. Last but not least, the gas prediction task belongs to the group of regression tasks while early detection of pump faults falls in the group of classification tasks.

The task of gas prediction regards timely prediction of day gas consumption that can remarkably help to the local distributor to keep in between the contracted consumption limits. Successful application of the instance-based learning to this task is represented by the distance weighted averaging model. In terms of this local model, the distance and kernel functions have to be defined. The presented solution applies the Euclidean distance metric with pre-processed (VDM proved to be a good transformation metric for nominal features), normalized and weighted features. The variable and the linear kernel functions are used. The performance bias batch optimizer using genetic

algorithms was employed to optimize all the model parameters. Issues of the optimal training set used for this optimization as well as for the model operation (prediction) were discussed and solved. The thesis studied applicability improvement of the distance weighted averaging method. This improvement concerns introduction of the correction weights. Idea of the method is based on local adaptation process utilizing difference between the queried instance and the selected neighbors. The target function values of the individual nearest neighbors are not weighted immediately but after correction driven by difference of the feature vectors of the queried instance and the given nearest neighbor.

The task of early detection of the pump faults focuses on construction of a robust decision tool transforming values of a set of on-line measured features into a distinct fault classification. The particular characteristic of the given task consists in existence of several classification scales. It means that the individual faults can be partially ordered into several scales of different severity of the given type of fault. Moreover, the process of the data acquisition gives a chance to use the additional knowledge (besides the original classification) when deciding about the competence of the given training instance into the given classification scale. Both these characteristics are taken into account in the new proposed classification algorithm - the competence estimate design method. The method fundamental is introduction of the competence estimate vectors instead of the original classifications for the training instances followed by application of the instance-based learning in order to find these vectors for unseen instances. The final step makes transformation of the competence estimate vectors into the final classifications of the unseen instances.