Summary of the doctoral dissertation

by Tomasz Orczyk, M.Sc., Eng.

Classification of incomplete data based on the classification committee

Thesis:

It is possible to maintain the accuracy of classification on incomplete data by selecting a committee of classifiers based on pre-selected features.

The purpose of the work was to develop a classification committee, designed to classify data in which there are features that do not have defined values. The classifier would be able to process incomplete feature vectors without the need to pre-fill them, and the classification would be based on pre-selected features. **Partial objectives** have been specified in the work:

1. Estimation of the impact of missing or removed features of the object on the quality of classification.
2. Developing the structure of the classification committee.
3. Selection of classifiers operating in the committee.
4. Developing a decision-making algorithm (fuser) for the classification committee.
5. Selection of distinctive features for individual object classes.
6. Testing the developed system on real data.
7. Verification of the usefulness of the developed classifier for the construction of the system for assessment of liver fibrosis in patients with hepatitis C based on the analysis of peripheral blood parameters.

The dissertation investigated the influence of the presence of *null* values in the data on the formation of incomplete reference (training) vectors depending on the size of the subspace of features on which the component classifiers of the committee work. The impact of the missing values on the quality of the classification has also been confirmed experimentally.

Based on the conclusions regarding the distribution of missing values of features among reference vectors, the structure of the classification committee was proposed, based on the division of feature space into one-element vectors.

For the proposed structure of the classification committee, a number of conventional classifiers were tested as component classifiers of the committee. As a component classifier of the committee being developed, the only classifier which benefited from such a committee structure, has been chosen – the *k*-NN classifier.

The Bayesian averaging, supplemented by the weighting factor for individual classes of reference objects, aimed at improving the quality of classification in relation to objects that are not very numerous in the reference set, has been proposed as the method of evaluating the classification committee decision.

A committee with such a structure performs initial, dynamic filtering of features, based on the vector of classified data. Features that do not have a defined value in this vector are ignored in the classification process. In order to improve the quality of classification, a method for pre-selection of features has been proposed, based on the component classifier of the proposed committee. This method uses a ranking of distinctive features for individual classes from the reference set, to indicate a suboptimal subset of the features on the basis of which the classification will be conducted.

The proposed SF*k*-NN/C committee classifier has been tested on a number of benchmark databases containing full real data. In order to determine the impact of the missing values of random features, in both classified and reference data, on the quality of classification, *null* values were artificially introduced into the data, replacing the existing values of randomly selected features. The tests were carried out without and with the initial selection of features.

Ultimately, the classifier was used to classify actual medical data - blood analysis of HCV infected patients. The undefined values in this data set occurred naturally. The test results were consistent with previously obtained results on data from which some values were artificially removed. Thus, the usefulness of the proposed SF*k*-NN/C classifier for the construction of the liver fibrosis assessment system in patients with hepatitis C has been confirmed.

The implementation of the partial objectives made it possible to confirm the thesis of the work. This confirmation is experimental, and supported by the results of statistical tests.