Description for the general public

Incompleteness of data can be characterised as a total lack of one or more attributes (features) describing the object (instance). The uncertainty is a more complex concept. The phenomenon has been studied for many years, and there are two main approaches to understanding the uncertainty of information – epistemic and ontic. In both of them an uncertain concept is described by a set of its possible representations. In the epistemic approach, without additional knowledge it is impossible to pick the right one among them, although the right one exists. In the ontic one, all representations are equally acceptable and there is therefore no need to distinguish between them. For example let us consider a concept of a monthly salary for assistant professor, described by the interval [$3000, $5000]. Such an interval would have an ontic interpretation if it described the minimum and maximum threshold of the salary for the position of assistant professor in some institution. On the other hand, if this interval was supposed to describe an actual salary of some particular assistant professor, it should be interpreted in an epistemic way.

In the past few years there has been a significant increase in interest in the problem of classification based on incomplete and uncertain data. This is mainly due increasingly widespread use of data mining techniques and machine learning in different areas of research, as well as everyday life. Therefore arises the need for processing of the data that has not been collected especially to be used in classification. Such data is often incomplete or inaccurate, and generally uncertain. Undoubtedly, uncertainty is widespread in real life and practical applications, and cannot be ignored.

Although there are many classification methods, few of them address the problem of classification based on uncertain data in general. The main disadvantage of the few existing methods is a high level of complexity, which prevents their use in decision support problems where high degree of transparency is needed. The proposed project aims to develop easy-to-interpret uncertain data classification methods.

It turns out that in the case of uncertain data, many known classification methods fail because they are not able to operate under incomplete information. The main scientific objective of the proposed project is to develop an effective classification method for uncertain data. During the research, pre-processing and classification algorithms of such data based on the similarity measures will be developed. Similarity measure for Hesitant Fuzzy Sets (HFS), which will enable adequate and effective modeling of widely understood incompleteness and uncertainty of data will be also defined.

Realisation of the project will contribute to a change in the approach to the problem of incomplete and uncertain data classification. Currently, the most commonly used methods are based on editing the data set, which consequently often lead to impoverishment (removing instances and/or attributes) or to simplification of the data structure (imputation).

The problem of data uncertainty is particularly important in medical diagnostics where incompleteness and uncertainty is often natural and permanent feature of the data. Hence, any attempt to neglect this problem leads to the models that do not reflect reality. Editing the data is generally not allowed in medical applications. Removing incomplete instances from often to small data set, significantly narrows the research possibilities. On the other hand, imputation based approach can not be used to classify specific medical case due to the high risk of wrong diagnosis because imputed values in no way reflect the current patient condition.

Results of the research will form the basis for the adoption of a completely different approach. Rather than ignoring or artificially improving the quality of the input data, it should be explicitly addressed in the data model, and then in the classifier during both the learning and testing. This approach offers many advantages comparing to the data editing. The classifier with more accurate (though incomplete) description of reality, will be able to learn the greater number of dependencies, and then make a better classification. It is also very important to be able to estimate how much particular classification result is uncertain. Of course, the data of very low quality (highly uncertain) can lead to an uncertain classification.

Certainly the development and popularisation of this approach to uncertain data classification will contribute its application in the problem of decision support, especially supporting medical diagnosis.