Application of explainable artificial intelligence algorithms in order to improve imbalanced data methods

The problem of the imbalanced data is often present in real life. Many times, the occurrence of infrequent events is of greater importance than that of more common appearances. In domains such as medicine, computer networks and bank security, instances of rare illness, computer attacks, or credit card frauds are crucial to be properly recognised, as their omission would have significant costs, both monetary or even in human life. The data imbalance, that is, a disproportion between the sizes of the classes, poses a great difficulty in a machine learning domain. When not considered, it may result in the rarer class, usually called the minority class, being recognised poorly or even being ignored at all. It is caused by the assumption for most classifiers that each miss-classification is of equal worth, which may lead to the bias to the class of the bigger size, called majority, as its more frequent selection leads to easier, smaller overall error. The problem applies not only to traditional machine learning algorithms but also to presently popular deep learning models. It is crucial to tackle it in some way, either on data or algorithmic levels.

The topic that has gained popularity in recent years is the explainability of artificial intelligence. It is related to the rapid growth of the usage of deep learning, in which models are black-box by nature. This, together with extending the diversity of applications in medicine, culture, banking and social life, established an urgent need to understand the way of working of the algorithms, reflected in, among others, the European Union Artificial Intelligence Act. As a result, many algorithms have been proposed that try to explain the decisions of the machine learning model. They aim to point out which of the problem's features had the most considerable influence on the final prediction. This knowledge may help with either selecting or improving the model. In the first case, an expert can examine whether the supposedly high predictive quality is not caused by some shortcuts or wrong assumptions. In the second, the information which features lead to learning wrong dependencies, like in the case of a horse being recognised by a footprint in the photo, allows the researchers to alter the data so that the classification algorithms can generalise and solve the actual problem better.

One of the goals of explainable artificial intelligence is to improve the quality of the models by learning their weaknesses and biases. It allows us to find problematic parts of probles attributes and alter them so the classification algorithm learns better. One of the biases it could counter is that of favouring one of the specific problem's classes, which often occurs for the imbalanced data problem. Moreover, frequently, it is not the disproportion of the classes that poses a challenge but their distribution in a feature space that could be recognised and conditioned. This is why, in this project, we form the following hypothesis:

It is possible to improve the quality of the imbalanced data classifier by employing explainable artificial intelligence techniques that gain knowledge about the reasons for the incorrect sample classification..

During this project, several methods employing explainable artificial intelligence algorithms will be developed. There will be proposed algorithms that depend on the models, forming the ensembles of classifiers and also model agnostic - creating an even dataset, allowing the usage of any classification method that suits the problem the best. The proposed methods will be analysed, evaluated and compared with the state-of-the-art imbalanced data algorithms. Then, they will be documented and shared in an open-source library.