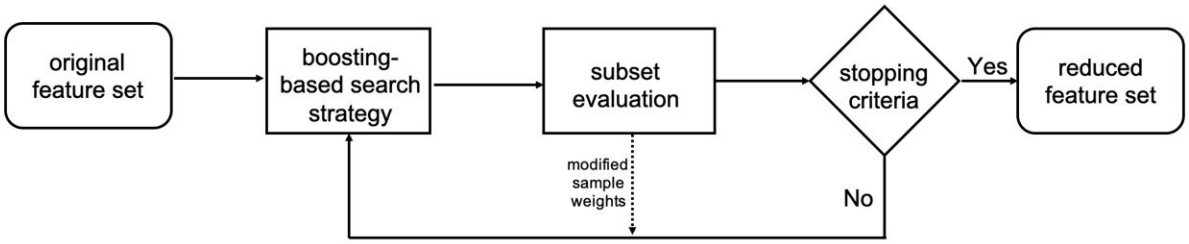
Finding the needles in a haystack of high-dimensional data sets

September 23 2021



How the newly developed feature selection algorithm works. Credit: University of Groningen

One of the challenges in the era of Big Data is dealing with many independent variables, also known as the "curse of dimensionality." Therefore, there is an urgent need to develop algorithms that can select subsets of features that are relevant and have high predictive powers. To address this issue, computer scientists at the University of Groningen developed a novel feature selection algorithm. The description and validation of their method was published in the journal *Expert Systems with Applications* on 16 September 2021.

The ability to select the smallest and most relevant subset of features is desirable for various reasons. First, it allows faster and, therefore, more scalable analysis. Second, it results in cheaper data acquisition and

storage. Third, it facilitates better explainability in the interaction between the selected features. "It is a misconception that the more features we add, the more information we have to make a better judgment," says George Azzopardi, assistant professor in Computer Science at the University of Groningen. "There are situations where some features may turn out to be completely irrelevant or redundant for the task at hand." Moreover, the task of explaining the outcome of a decision that is made by a computer algorithm becomes more complicated with an increasing number of independent variables.

Interactions

"Feature selection is widely used and it is achieved using varying approaches," says Ahmad Alsahaf, a postdoctoral researcher at the UMCG and the first author of the paper. Identifying the right features is quite challenging, it is like finding a needle in a haystack. A naive approach to select the best subset would be a brute force selection that evaluates all possible combinations of features. "However, this approach is intractable for large numbers of features," says Alsahaf. Other approaches use, for example, statistical methods to measure the significance of each individual feature with respect to the dependent variable.

Azzopardi explains that "While such approaches are very fast, they do not consider the possible interaction between the independent variables. For instance, while two independent variables may have very low discriminative powers when considered individually, they may have very strong predictive powers when considered together." Alsahaf added that "a common example is the interaction of epistatic genes, where the presence of one gene affects the expression of another. Feature selection algorithms must be able to detect such interactions."

Boosting

The computer scientists designed a novel feature selection algorithm that relies on what is known as boosting, which they called FeatBoost. Alsahaf says that they "use a decision tree-based model to select the most relevant features. We subsequently create and evaluate a classification model using the selected features so far. Any samples that are wrongly classified will be given more emphasis in determining the next set of most relevant features, a process called boosting. These steps are repeated until the performance of the classification model cannot improve any further."

In the paper, the scientists demonstrate the effectiveness of their algorithm on various benchmark data sets with different properties and show how it outperforms other well-known methods, such as Boruta and ReliefF. In particular, they claim that their algorithm achieves higher accuracies with fewer features on most of the data sets that they used for evaluation.

The source code for the algorithm is available on GitHub.

More information: Ahmad Alsahaf et al, A framework for feature selection through boosting, *Expert Systems with Applications* (2021). DOI: 10.1016/j.eswa.2021.115895

Source code: github.com/amjams/FeatBoost

Provided by University of Groningen

https://techxplore.com/news/2021-09-needles-haystack-high-dimensional.html