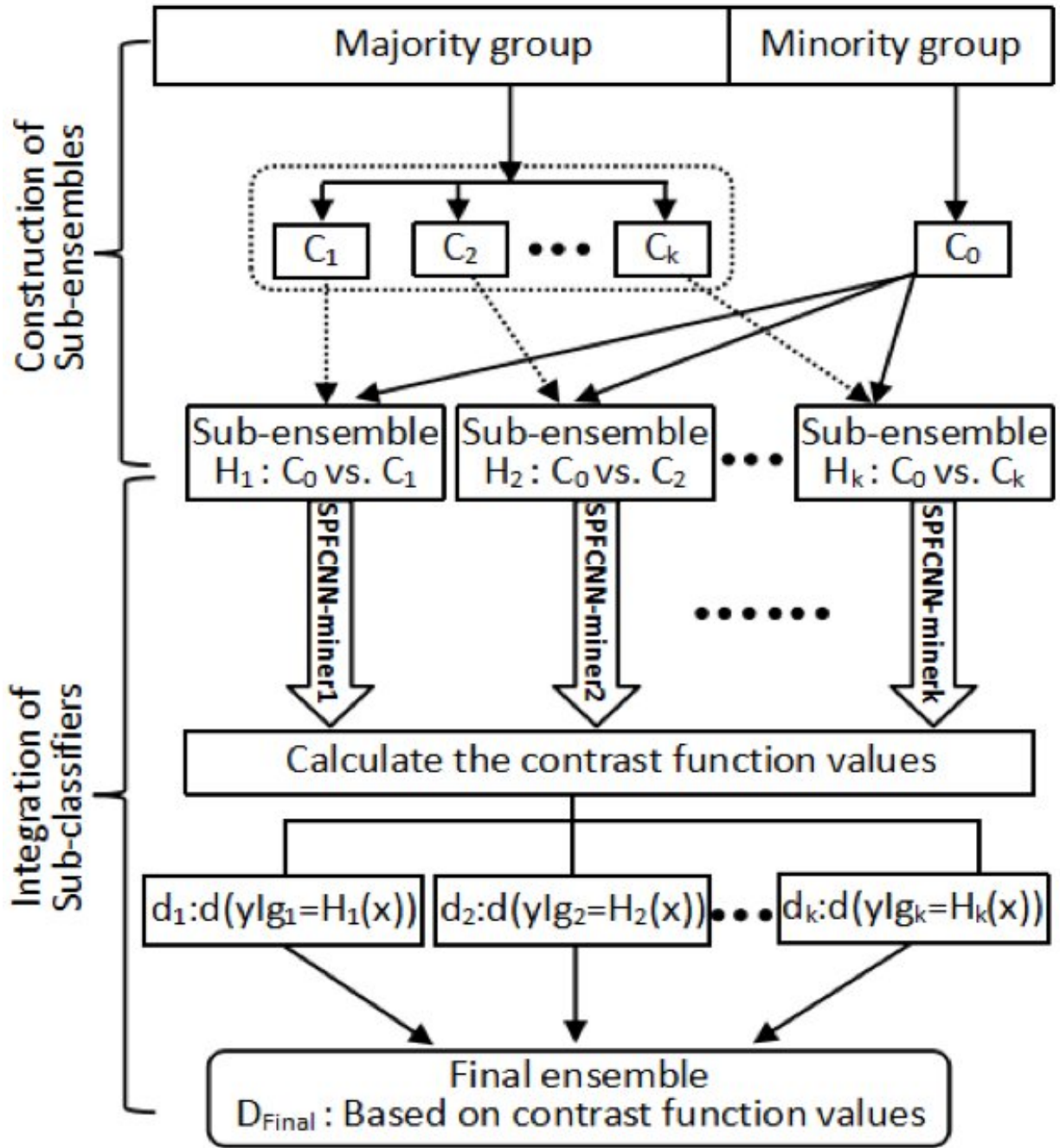# SPFCNN-Miner: A new classifier to tackle class-unbalanced data

June 19 2019, by Ingrid Fadelli

The flowchart if MLF. Credit: Zhao et al.

Researchers at Chongqing University in China have recently developed a

cost-sensitive meta-learning classifier that can be used when the training data available is high-dimensional or limited. Their classifier, called SPFCNN-Miner, was presented in a paper published in Elsevier's *Future Generation Computer Systems*.

Although machine learning classifiers have proved to be effective in a variety of tasks, to achieve optimal results, they often require a vast amount of training data. When data is high-dimensional, limited or unbalanced, most classification methods are unable to achieve a satisfying performance. In their study, the team of researchers at Chongqing University set out to better understand these data-related challenges and develop a classifier that can overcome them.

"We used Siamese networks that are suitable for few-shot learning where a little data is available to learn high-dimensional and limited data, and apply the idea of combining 'shallow' and 'deep' approaches to design parallel Siamese networks that can better extract simple or complex features from a variety of datasets," Linchang Zhao, one of the researchers who carried out the study, told TechXplore. "The main objectives of our study were to solve the data class-imbalanced problem and get the best possible classification results on such datasets."

Zhao and his colleagues developed a Siamese parallel fully-connected neural network (SPFCNN) and applied it to problems with class-unbalanced data distributions. To transform their cost-insensitive SPFCNN into a cost-sensitive approach, they used a technique called 'cost-sensitive learning."

First, the researchers divided the majority group in a dataset based on inner-product transformed features. This ensured that the size of each sub-group in a majority group was close to that of the minority group. In addition, they structured some sub-ensembles using the minority group vs. each partition obtained.

"Next, we applied n SPFCNN-miners to all sub-ensembles, each sample point x_j can be expressed by its corresponding measures$(d_{j1},\ldots, d_{jn})$, each sub-classifier can be transformed into a measure of contrastive loss function by fitting the SPFCNN," Zhao explained. "Finally, n SPFCNN-miners were integrated as a final classifier according to the values of contrastive function."

The approach devised by Zhao and his colleagues has numerous advantages that set it apart from other classifiers. First, their Meta-Learner Function (MLF) can be used to partition the majority group in a dataset based on the inner-product transformed features, which results in the transformed data containing information related to distances and angles between items in the minority and majority groups.

"The angles between the majority group and the minority group can be viewed as the expression of related locations and then represent the related direction of the majority group to the minority group," Zhao explained.

A further advantage of the new SPFCNN-Miner classifier is that, like other Siamese networks, it can effectively extract the highest-level features from a small amount of samples for few-shot learning. Moreover, parallel Siamese networks are designed to adaptively learn simple or complex features from different dimensions of data attributes.

Zhao and his colleagues evaluated their approach in a series of computational tests, using both cost-insensitive and cost-sensitive versions of the SPFCNN classifier. They found that the cost-sensitive approach outperformed all of the classifiers they compared it with.

"The experimental results show that our SPFCNN is a competitive approach and is able to improve the classification performance more significantly compared with the benchmarked approaches," Zhao said.

"We found that the performance of our model did not improve as the sample size increased, but was greatly affected by the imbalance rate. The performance obtained by incorporating the cost-sensitive learning into our model is more stable."

The study carried out by Zhao and his colleagues introduces a new method that could be used by researchers to enhance the performance of classifiers when data is limited or unbalanced. In addition, their findings suggest that balancing the number of positive and negative samples can be more effective than generating a larger number of artificial samples. For instance, their approach can integrate different misclassification costs as it completes a classification task, which makes it more robust than other techniques used to address issues unbalanced data-related issues.

"In the future, we plan to use techniques such as random walk matrices, circulant weight sharing and Huffman coding to compress our model, and the loosely-connected technology or paralleled pruning-quantization method will be used to lightweight the proposed SPFCNN model," Zhao said.

Linchang Zhao ORCID ID: orcid.org/0000-0002-6502-587X