|  |  |
| --- | --- |
| Reading note: γk-nn model for imbalanced data  Celian RINGWALD  https://github.com/datalogism/AdjustedNearestNeighborAlg  [This Photo](https://pngimg.com/download/73412) by Unknown Author is licensed under [CC BY-NC](https://creativecommons.org/licenses/by-nc/3.0/) |  |

Paper

Viola, Rémi and Emonet, Rémi and Habrard, Amaury and Metzler, Guillaume and Riou, Sébastien and Sebban, Marc, 'An Adjusted Nearest Neighbor Algorithm Maximizing the F-Measure from Imbalanced Data', International Conference on Tools with Artificial Intelligence (ICTAI), 2019 <https://arxiv.org/abs/1909.00693/>

Summary

The current scientific paper is extending Nearest-Neighbors (k-NN model) for detecting anomalies in imbalanced data. This machine learning method get the k closest point of a query depending to chosen distance (typically the Euclidean one), and assign a class based on a majority vote-based decision. This simple method has very high capacity on learning non-linear relations but need to be adapted in context of weakly represented data. Borrowing tools from sampling, metric learning, and weighted distance-based strategies, the proposed **γk-nn** model evaluate the best γ weight, that will rescale the decision area around the positive examples. They first compare this approach on a large panel of imbalanced dataset and compare it with state-of-art models, they also demonstrate the effective complementarity of the strategy with sampling methods. The results of these experiments highlight the performance of the γk-nn in term the F-Measure and describe the impact of the imbalanced ratio on it.

Distance-based, metric learning and sampling strategies

Concerning strategies spotlighted in the paper, the first one are distance-based methods. The most basic one is **wk-NN** (Dudani, 1976) consist to associate to each neighbors a weight depending to the distance between them and the given query point. Another line consists to estimate the local sparcity around minority examples and taking it in account by adjusting posterior probabilities as in the **kRNN** model (Zhang, et al. 2017). The **cwk-NN** model is proposing to weight directly the Euclidean distance by class (Barandela et al 2003), this method is extended by the **γk-nn** model.

The second strategy is linked to metric learning method and optimization, another way where this current paperwork is projected too via the R. Viola Thesis. It consists to learn under constraints parameters linked to a distance, for exemple the Mahalanobis one in the LMNN (Weinberger 2009) and in the ITML (Davis et Al. 2007), only focused on positive examples. This is how will be fixed the **γ parameter of the presented model.**

**Finally**

The model

Results

Results