Reading note: γk-nn model for imbalanced data

Celian RINGWALD

https://github.com/datalogism/AdjustedNearestNeighborAlg

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 gets the k closest point of a query depending on chosen distance (typically the Euclidean one) and assigns a class based on a majority vote-based decision. This simple method has a very high capacity for learning non-linear relations but needs to be adapted in the context of weakly represented data. Borrowing tools from sampling, metric learning, and weighted distance-based strategies, the proposed **γk-nn** model evaluates the best g weight, which will rescale the decision area around the positive examples. They first compare this approach on a large panel of imbalanced datasets 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 terms of 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 is **distance-based methods**. The most basic one is *wk-NN* (Dudani, 1976) consists to associate each neighbor a weight depending on the distance between them and the given query point. Another line consists to estimate the local sparsity around minority examples and taking it into account by adjusting posterior probabilities as in the *kRNN* model (Zhang, et al. 2017). The *cwk-NN* model is proposing to weighting 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**s and optimization, another way where this current paperwork is projected today notably by the R. Viola Thesis. It consists to learn under constraints parameters linked to a distance, for example the Mahalanobis one in the *LMNN* (Weinberger 2009) and 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 paper presents the advantages of **sampling methods**, more precisely methods that compensate artificially the lack of positive example. Based mostly on the *Synthetic Minority Over-sampling Technique* (Chawla et al. 2002), they create random and synthetic on the line between a given point his neighbors. An iterative process is done when chosen balanced rate is obtained. *Border-line Smote* (Han et al. 2005) extend it by focusing only on points that have more negative than positives in neighbors. Concerning over-sampling methods article present also the *ADASYN* model (He et al. 2005). But it also compares them to strategies that combine oversampling with under-sampling, as the *ENN* (Wilson, 1972), built on the top of the *SMOTE*, and *the Tomek's link* (Tomek 1976).

The model

In the presented landscape designed by the paper, the model is **defining the distance** as following, where $S_{-}$are the negatives data points and $S_{+}$are the positives.

$$d_{\gamma}(x, x_i) = \begin{cases} d(x, x_i), & if\ x_i \in S_{-} \\ \gamma . d(x, x_i), & if\ x_i \in S_{+} \end{cases}$$

Where the γ parameter is **learned by maximizing** the non-convex F-measure, apply only on positives examples comparisons. The paper also describes the consequences of $\gamma \leq 1$ and $\gamma > 1$ cases on the False Negative probability and the False Positive probability. They underline the $\gamma \leq 1$ case speed-up the convergence of the algorithm.
This one has the same overall complexity that the k-nn. He first finds the nearest positives and negatives examples to a given point, computes their distance to him, and weights the positives one. Then it uses the k closest one to make the classification decision.

Results of the experiments

The paper first compares the γk-nn model with the previously presented distance-based and metric learning methods on the large panel of imbalanced datasets with and without using sampling strategies. The results underline the effectiveness of the γk-nn and present also the LMNN as a competitive challenger. Concerning the imbalanced ratio, the experiments conclude to relative robustness of the γk-nn. They also show the complementarity of the presented method with oversampling no matter the choice of strategy. Moreover, the registration of the project with the General Directorate of Public Finance has allowed us to illustrate the efficiency of the model on real data.

Connections with Uncertain complex graph

We could see some connections between this paper and current research projects of the Hubert Curien laboratory, notably with the “*Complex Graph Analysis for the Detection of Corruption in Public Procurement”* thesis project. In another hand k-nn is also applied in different bibliographic sources of the “*Complex data analysis with incomplete or uncertain data*” : with shortest path for classification (Pfeiffer et al. 2011), or in Kassiano et al. (2017) article in the nearest-neighbor searching task. We signals also that Uncertainty could also be faced with imbalanced data, even more, when we work with real data.