Proposition de stage de recherche - Année 2021/2022

Subject : Handling classes' imbalance in supervised classification for medical diagnostics

Lieu : LAMADE - Pôle Sciences des Données - Université Paris Dauphine – PSL
Place du Maréchal de Lattre de Tassigny - 75775 PARIS Cedex 16
Contact : Sana Ben Hamida (sana.mrabet@dauphine.psl.eu)

Objective: Study and compare three different approaches to handle classes' imbalance in medical data: data pre-processing with over/under sampling, synthetic minority over-sampling and active sampling.

Description: The classification of highly imbalanced data is a big challenge for machine learning techniques. To deal with this challenge, many solutions have been proposed that could be classified in three categories: data pre-processing with under/oversampling technique that creates a training sample with a new instances distribution [2], active sampling that changes the training sampling throw the learning process [3], and the Synthetic Minority Over-sampling Technique (SMOTE) [1] that creates new synthetic instances in the minority class. The efficiency of each approach depends on the context. For the medical diagnostics, if the input data contains categorical attributes, the SMOTE methods could be not suitable [4]. Otherwise, if the data imbalance ratio is high, using the under/oversampling could induce loss of information in the training sample.

The purpose of this work is to compare the performance of three approaches of some given medical data applied with an evolutionary classification technique. This study aims to give some guidelines to decide the appropriate imbalance handling method according to data characteristics. Otherwise, a new solution could be proposed by combining the three tested approaches. The idea of second part of this work is to generalise the ability to deal with the classes' imbalance in different machine learning contexts by an automatic selection of the appropriate approach.

Data and material: Four medical data sets are selected from the machine learning archive http://archive.ics.uci.edu/ml/datasets.php (other data sets could be proposed). All the solutions will be implemented in Python using some ML libraries such as DEAP for evolutionary classification and scikit-learn for data pre-processing, and other codes implementing SMOTE that have been made available by their authors on https://github.com/.

Some references

[1] D. Elreedy, A. F. Atiya, A Comprehensive Analysis of Syntheic Minority Oversampling Technique (SMOTE) for handling class imbalance, Information Science 505 (2019) 32-64.

[2] Hunt, R., Johnston, M., Browne, W., Zhang, M.: Sampling methods in genetic programming for classification with unbalanced data. In: Li, J. (ed.) AI 2010. LNCS (LNAI), vol. 6464, pp. 273–282. Springer, Heidelberg (2010).

[3] Hmida H. Ben Hamida S., Borgi A. Rukoz M.. Sampling Methods in Genetic Programming Learners from Large Datasets : A Comparative Study. Volume 529 of Advances in Intelligent Systems and Computing, pages 50–60, 2016. (Cité en pages 5 et 6.)

[4] Z. Xu, D. Shen, T. Nie, Y. Kou, N. Yin, X. Han, A cluster-based oversampling algorithm combining SMOTE and k-means for imbalanced medical data, Information Sciences, 572 (2021) 574-589

Profil du candidat :

L'offre s'adresse à un étudiant en M2 ou équivalent en Informatique, avec bonnes connaissances en Machine Learning et en programmation Python.
Durée:5-6 mois