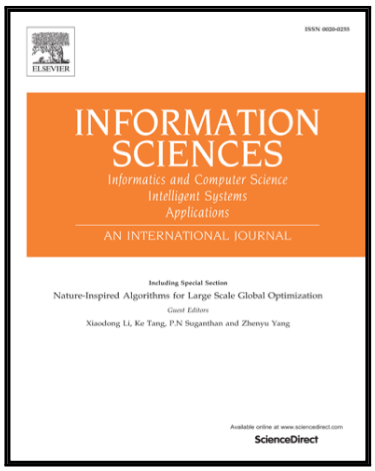

The study of under- and over-sampling methods' utility in analysis of highly imbalanced data on osteoporosis

M. Bach, A. Werner, J. Żywiec, W. Pluskiewicz

PII: S0020-0255(16)30895-7
DOI: 10.1016/j.ins.2016.09.038
Reference: INS 12534

To appear in: *Information Sciences*

Received date: 28 October 2015
Revised date: 6 September 2016
Accepted date: 14 September 2016

Please cite this article as: M. Bach, A. Werner, J. Żywiec, W. Pluskiewicz, The study of under- and over-sampling methods' utility in analysis of highly imbalanced data on osteoporosis, *Information Sciences* (2016), doi: 10.1016/j.ins.2016.09.038

The study of under- and over-sampling methods' utility in analysis of highly imbalanced data on osteoporosis

M. Bach[a], A. Werner[a], J. Żywiec[b], W. Pluskiewicz[c]

[a] *Silesian University of Technology, Gliwice, Poland*
[b] *Medical University of Silesia, Katowice, Poland*
[c] *Department and Clinic of Internal Diseases, Diabetology and Nephrology, Metabolic Bone Diseases Unit, Medical University of Silesia, Katowice, Poland*

Abstract

Osteoporosis is a frequent bone disease without typical early symptoms but with serious complications e.g. low-energy bone fractures. Patients with risk factors should be screened for proper diagnosis as early as possible. Unfortunately, the registered medical data are often highly imbalanced. That is why the machine-based data processing is difficult or even impossible. Considering this, our goal was to search for the best method of coping with the problem of imbalancing in relation to the analysed data regarding the osteoporotic patients. Therefore, we checked several paradigms of classifiers in synergy with preprocessing techniques to address the inner skewed class distribution of the data.

In the source dataset 92.6% of instances related to patients without any fractures (negative cases) and only 7.41% to patients (positive cases) who reported at least one fracture. To alleviate class imbalance there were examined not only data-level methods which in fact modify the input dataset, but also ensemble ones that strengthen the results of the base algorithms. In the first group the under- and over-sampling methods were used, such as random undersampling, edited nearest neighbours and synthetic minority over-sampling techniques, while in the second one – a range of methods based on various subsets of training data were analysed. Also various combinations

Email addresses: malgorzata.bach@polsl.pl (M. Bach), aleksandra.werner@polsl.pl (A. Werner), jzywiec@sum.edu.pl (J. Żywiec), osteolesna@poczta.onet.pl (W. Pluskiewicz)

URL: http://www.polsl.pl (M. Bach), http://www.polsl.pl (A. Werner), http://www.sum.edu.pl/ (J. Żywiec), http://www.sum.edu.pl/ (W. Pluskiewicz)