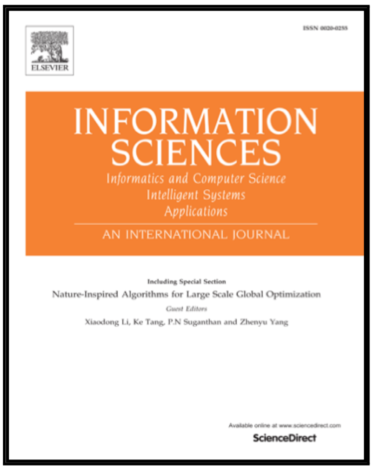

ISSN 0020-0255

ELSEVIER

INFORMATION SCIENCES

Informatics and Computer Science
Intelligent Systems
Applications

AN INTERNATIONAL JOURNAL

Including Special Section
Nature-Inspired Algorithms for Large Scale Global Optimization
Guest Editors
Xiaodong Li, Ke Tang, P.N Suganthan and Zhenyu Yang

Available online at www.sciencedirect.com
ScienceDirect

A Genetic Algorithm Approach to Optimising Random Forests Applied to Class Engineered Data

Eyad Elyan, Mohamed Medhat Gaber

PII: S0020-0255(16)30578-3
DOI: 10.1016/j.ins.2016.08.007
Reference: INS 12410

To appear in: *Information Sciences*

Received date: 30 October 2015
Revised date: 25 July 2016
Accepted date: 3 August 2016

Please cite this article as: Eyad Elyan, Mohamed Medhat Gaber, A Genetic Algorithm Approach to Optimising Random Forests Applied to Class Engineered Data, *Information Sciences* (2016), doi: 10.1016/j.ins.2016.08.007

A Genetic Algorithm Approach to Optimising Random Forests Applied to Class Engineered Data

Eyad Elyan, Mohamed Medhat Gaber

School of Computing Science and Digital Media
Robert Gordon University
Riverside East, Garthdee Road, Aberdeen
AB10 7GJ, UK
Email: {e.elyan,m.gaber1}@rgu.ac.uk

Abstract

In numerous applications and especially in the life science domain, examples are labelled at a higher level of granularity. For example, binary classification is dominant in many of these datasets, with the positive class denoting the existence of a particular disease in medical diagnosis applications. Such labelling does not depict the reality of having different categories of the same disease; a fact evidenced in the continuous research in root causes and variations of symptoms in a number of diseases. In a quest to enhance such diagnosis, datasests were decomposed using clustering of each class to reveal hidden categories. We then apply the widely adopted ensemble classification technique Random Forests. Such class decomposition has two advantages: (1) diversification of the input that enhances the ensemble classification; and (2) improving class separability, easing the follow-up classification process. However, to be able to apply Random Forests on such class decomposed data, three main parameters need to be set: number of trees forming the ensemble, number of features to split on at each node, and a vector representing the number of clusters in each class. The large search space for tuning these parameters has motivated the use of Genetic Algorithm to optimise the solution. A thorough experimental study on 22 real datasets was conducted, predominantly in a variety of life science applications. To prove the applicability of the method to other areas of application, the proposed method was tested on a number of datasets from other