Expert Systems with Applications 40 (2013) 4637–4647

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Data weighting method on the basis of binary encoded output to solve multi-class pattern classification problems

Kemal Polat *

Abant İzzet Baysal University, Faculty of Engineering and Architecture, Department of Electrical and Electronics Engineering, Bolu, Turkey

ARTICLE INFO

Keywords:
Data weighting
Binary encoded output based data weighting (BEOBDW)
k-NN classifier
Multi-class data classification

ABSTRACT

Data weighting is of paramount importance with respect to classification performance in pattern recognition applications. In this paper, the output labels of datasets have been encoded using binary codes (numbers) and by this way provided a novel data weighting method called binary encoded output based data weighting (BEOBDW). In the proposed data weighting method, first of all, the output labels of datasets have been encoded with binary codes and then obtained two encoded output labels. Depending to these encoded outputs, the data points in datasets have been weighted using the relationships between features of datasets and two encoded output labels. To generalize the proposed data weighting method, five datasets have been used. These datasets are chain link (2 classes), two spiral (2 classes), iris (3 classes), wine (3 classes), and dermatology (6 classes). After applied BEOBDW to five datasets, the *k*-NN (nearest neighbor) classifier has been used to classify the weighted datasets. A set of experiments on used real world datasets demonstrated that the proposed data weighting method is a very efficient and has robust discrimination ability in the classification of datasets. BEOBDW method could be confidently used before many classification algorithms.

1. Introduction

Data preprocessing methods are widely used in pattern recognition and machine learning applications. In order to produce the good and robust results in classification or prediction applications, data preprocessing methods including data cleaning, data transformation, data projection etc. The causes of using data preprocessing methods prior to classifier algorithms are to clean out noisy data, to impute missing values, and to remove redundant data from datasets. In the data cleaning, the problems to be encountered are missing values, inconsistencies, and normalization. In the data transformation, the conducted problems are data reduction, feature selection, and dimensionality reduction. To simplify the datasets, clustering algorithms and data visualization methods are used before classifier algorithms (Kotsiantis, Kanellopoulos, & Pintelas, 2006).

In this paper, as data preprocessing, data weighting method called binary encoded output based data weighting (BEOBDW) has been proposed both to increase the classification performance and to simplify the complex of datasets. Also, the other aim of BEOBDW is to transform from non-linearly separable datasets to linearly separable datasets.

To show the applicability of proposed BEOBDW method, multi-class datasets have been used. In literature, many studies have been conducted in solution of multi class classification problems. Among these studies, Shiraishi et al. proposed a new approach for combining binary classifiers. Their method trains a combining method of binary classifiers with statistical techniques such as penalized logistic regression, stacking, and a sparsity promoting penalty (Shiraishi & Fukumizu, 2011). A dynamic over-sampling procedure has been proposed to improve the previous classification of imbalanced datasets with more than two previous classes by Fernández-Navarro et al., 2011. Their procedure has been incorporated into a memetic algorithm (MA), which optimizes radial basis functions neural networks. They applied their method to multi-class classification problems (Fernández-Navarro, Hervás-Martínez, & Gutiérrez, 2011). Jianhua Xu improved traditional binary support vector machine by introducing an approximate ranking loss as its empirical loss term for building a novel support vector machine for multi-label classification (Xu, in press). Brucker et al. investigated hierarchy extraction from results of multi-label classification (MC). Their system classifies multi-label data and extracts a class hierarchy from multi-label predictions. It is established upon a combination of a novel multi-label extension of the fuzzy Adaptive Resonance Associative Map neural network with an association rule learner (Brucker, Benites, & Sapozhnikova, 2011). A formal analysis of the consistency of pairwise bipartite rankings by uncovering the conditions

* Tel.: +90 530 561 9226; fax: +90 374 253 45 58.
E-mail address: kemal_polat2003@yahoo.com

0957-4174/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2013.02.004

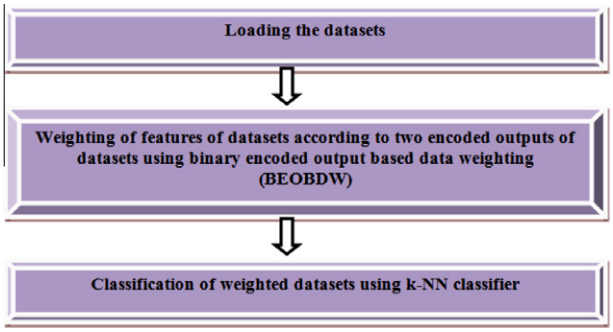

Fig. 1. The flowchart of proposed hybrid method.

Table 1
The binary codes of outputs of a dataset.

Number of class in dataset	1. Output	2. Output
1	2^0	2^1
2	2^2	2^3
3	2^4	2^5
4	2^6	2^7
5	2^8	2^9
...	...	...
...	...	...
n	2^{2n-2}	2^{2n-1}

Table 2
An example dataset explaining the working of OBEBFW.

Features			Binary coded outputs	
A_1	A_2	A_3	O_1 ($2^0 = 1$)	O_2 ($2^1 = 2$)

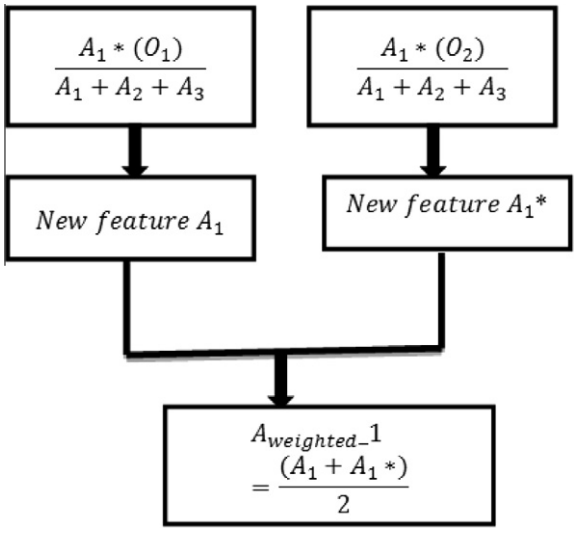

Fig. 2. A weighting process for first dimensional of dataset using binary encoded output based data weighting.

under which they can be equivalently expressed in connection with a single ranking has been conducted by Waegeman and De Baets (2010). Galar et al. used different base classifiers so as to observe the suitability and potential of each combination within each classifier. They compared the performance of these ensemble techniques with the classifiers' themselves (Galar, Fernández, Barrenechea, Bustince, & Herrera, 2011).

Apart from the literature studies, a novel data preprocessing (weighting) method has been proposed and called binary encoded

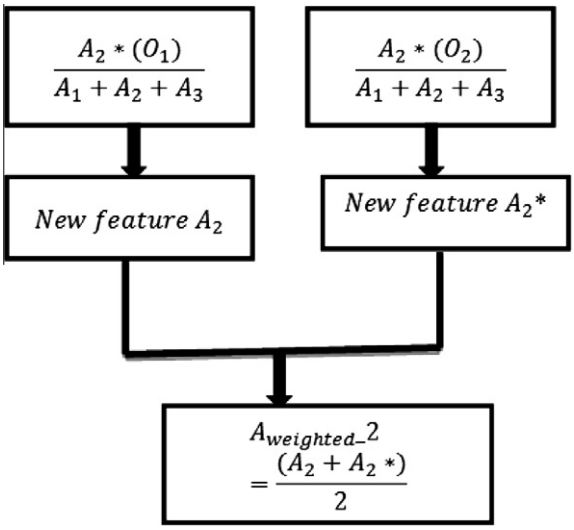

Fig. 3. A weighting process for second dimensional of dataset using binary encoded output based data weighting.

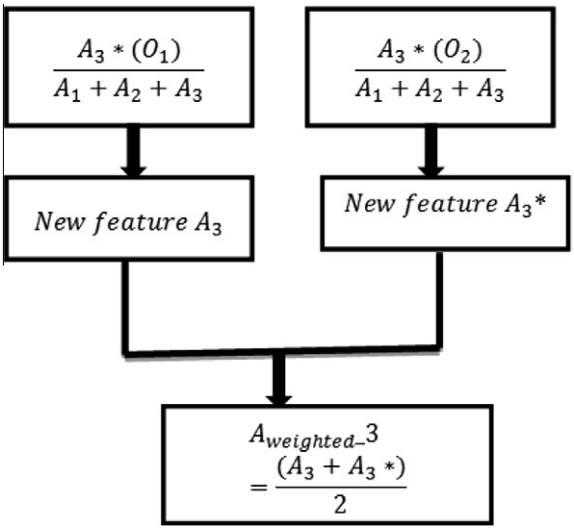

Fig. 4. A weighting process for third dimensional of dataset using binary encoded output based data weighting.

output based data weighting (BEOBDW). This preprocessing method consists of two stages. Firstly, the output labels of dataset used has been encoded using binary numbers and by this way two encoded outputs have been obtained. Secondly, the dataset has been weighted using the relationships between features of dataset and two encoded outputs. After BEOBDW applied to datasets, *k*-NN classifier has been used. To select the best *k* value in *k*-NN classifier in the classification of datasets, various *k* values have been used and selected the best k value in the classification of datasets. The used datasets to evaluate the proposed method are chain link (2 classes), two spiral (2 classes), iris (3 classes), wine (3 classes), and dermatology (6 classes). Iris, wine, and dermatology datasets have been taken from UCI (University of California, Irvine) machine learning database (UCI machine learning database., 2013). To test the proposed method combining BEOBDW and *k*-NN classifier, the classification accuracy, recall, TNR (true negative rate), precision, g-mean 1, g-mean 2, f-measure values have been used.

The remainder of this paper is organized as follows. Section 2 gives the method. Section 3 presents the results and discussion. Finally, Section 4 concludes the obtained results and findings.