Information Sciences 000 (2016) 1–15

Contents lists available at ScienceDirect

Information Sciences

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

Medical decision support system for extremely imbalanced datasets

Swati Shilaskar*, Ashok Ghatol[1], Prashant Chatur[2]

Government College of Engineering, Amravati, India

ARTICLE INFO

Article history:
Received 29 October 2015
Revised 2 August 2016
Accepted 21 August 2016
Available online xxx

Keywords:
Imbalanced dataset
Evolutionary algorithm
Medical diagnosis
Particle swarm
Physiological parameters
Synthetic sampling

ABSTRACT

Advanced biomedical instruments and data acquisition techniques generate large amount of physiological data. For accurate diagnosis of related pathology, it has become necessary to develop new methods for analyzing and understanding this data. Clinical decision support systems are designed to provide real time guidance to healthcare experts. These are evolving as an alternate strategy to increase the exactness of diagnostic testing. Generalization ability of these systems is governed by the characteristics of dataset used during its development. It is observed that sub pathologies have a much varied ratio of occurrence in the population, making the dataset extremely imbalanced. This problem can be resolved at both levels i.e. at data level as well as algorithmic level. This work proposes a synthetic sampling technique to balance dataset along with Modified Particle Swarm Optimization (M-PSO) technique. A comparative study of multiclass support vector machine (SVM) classifier optimization algorithm based on grid selection (GSVM), hybrid feature selection (SVMFS), genetic algorithm (GA) and M-PSO is presented in this work. Empirical analysis of five machine learning algorithms demonstrate that M-PSO statistically outperforms the others.

1. Introduction

Collection of either measured or observed parameters of patients during examination and respective diagnosis of the sample by healthcare expert constitutes a medical dataset. Its sample count depends on the number of patients visiting a clinic. It is likely that the number of samples generated for various pathologies have varied sample count which makes the dataset imbalanced. Evaluating imbalanced datasets is a very important problem from performance and algorithmic perspective. If the classification categories are not represented in an appropriate (nearly equal) proportion or if there are significantly more data points of one class and fewer occurrences of the other class, then the dataset is said to be imbalanced. A bias is developed towards majority due to non-uniform distribution of samples. Many real world problems are characterized by imbalanced data. For pathology detection, minority class is of potential interest as it indicates disease diagnosis capability of the system. In such cases there is a higher cost for misclassifying the minority class than misclassifying

* Corresponding author, Assistant Professor, Vishwakarma Institute of Technology, Bibwewadi, Pune 411037, India.
E-mail addresses: shilaskar_swati@rediffmail.com, swati.shilaskar@vit.edu (S. Shilaskar), vc_2005@rediffmail.com (A. Ghatol), chatur.prashant@gcoea.ac.in (P. Chatur).
[1] Former Vice Chancellor, Dr. Babasaheb Ambedkar Technological University, Lonere, India.
[2] Associate Professor and Head, Department of Computer science and Engineering.

http://dx.doi.org/10.1016/j.ins.2016.08.077
0020-0255/

the majority class. Our work has focused on classification of multiclass imbalanced datasets from medical domain. In this paper the topics are arranged as follows:

Literature survey carried out for data conditioning with synthetic sampling techniques is given in Section 2. Description of imbalance and dimensions of datasets is given in Section 3. It also includes comparative study of classification performance achieved by various algorithms for benchmark datasets in literature. Data conditioning scheme is described in Section 4. Experiments set up for multiclass classifiers and graphical illustrations of observations noted during optimization process are given in Section 5. Section 6 consists of performance evaluation methods for multiclass classification and results.

2. Literature survey

Many machine learning approaches are developed to deal with the imbalanced data. Either classifier algorithm is designed to handle the bias or data is externally synthesized and dataset is modified to be a balanced dataset. The oversampling of minority classes gives more accurate results than under sampling of majority classes. Under sampling by deleting weak discriminating samples away from hyper plane and oversampling by generating virtual samples in the proximity of hyper plane of SVM is suggested in [16]. Synthetic sampling technique for binary classification using fivefold cross validation technique is used in [31]. B. Das et al. [12] suggested one class classifier approach, making 'sensitivity' - a key evaluating parameter for imbalanced data. B. Almogahed et al. [2] used a semi-supervised learning method to identify the most relevant instances to establish a well-defined training set. According to J. F. Diez-Pastor [15] and S. Barua et al. [8], diversity increasing techniques for synthetic samples improve classification performance. Literature indicates absence of a fixed rule to find the correct distribution of samples for a learning algorithm [11]. X. Wan et al. [46] used difference between majority and minority class samples used to define cost function without priori knowledge. Literature suggests elimination of samples which do not constitute support vectors of the classifier. For a multiclass classification system this technique may not be appropriate, as a sample insignificant for one, might be vital for classifying other pathology. For balancing the dataset, under sampling, oversampling or a combination of both can be used. Under sampling may potentially remove certain important examples leading to over fitting. Oversampling on the other hand introduces additional computational task. The class imbalance and the nature of the learning algorithm both are strongly related. Artificially generated instances are not real data hence their true class may be questioned. There are two types of feature vectors,

(1) Directly collected e.g. symptoms, age, temperature etc.
(2) Extracted by signal processing techniques e.g. features extracted from speech, ECG, EEG, EMG signal, medical image etc.

It is appropriate to rely on domain experts' opinion about true class of synthetic samples in the first type. In the second type, where features are transformed, one may need to rely on the data itself [40].

In this work, we propose a technique to balance the dataset with minimal possible compromise to the diversity of feature space. We use multiclass medical datasets with varied degree of imbalance. Classifier optimization is carried out using cross validation technique [28] The performance measures for multiclass classification need to be selected carefully. Evaluation of classifiers is based on the performance for reserved samples which were not involved in the training process.

3. Dataset description and related work

We employed pathological multiclass imbalanced datasets for analysis. The datasets were chosen on the basis of their class distributions and sizes. Various physiological parameters collected during observation and treatments are parts of medical dataset. The performance of various classification methods in literature, implemented on relevant datasets is given below.

Vani Dataset: 'Vani' is a word of Sanskrit origin, meaning 'voice' . This database is developed at Vani speech therapy center, Jabalpur, India. These are 16-bit resolution speech samples with sampling rate of 22,050 Hz. The samples are sustained phonation of /a/ recordings from patients with a variety of vocal cord pathologies including organic, neurological and traumatic disorders. The data set contains 11 samples with mild pathology, 19 with severe pathology and 93 healthy speakers' samples. Pathological class is categorized after expert's opinion about the perceived pathology. Expert medical practitioner's opinion and suggestions from previous study [5] were applied for categorization of samples. Healthy voice samples are categorized in the following ways:

- The subjective feeling that the speaker has no perceived laryngeal pathology.
- An adequate voice for the age, gender and cultural group of the speaker.
- An adequate pitch, tone, volume and flexibility of diction.
- No history of surgery related with any laryngeal pathology.

22 features based on voice quality are extracted from speech samples from this dataset. The features are based on fundamental frequency, harmonic to noise ratio (HNR), normalized noise energy (NNE) and glottal-to-noise excitation ratio (GNE) features in combined feature set for evaluation. For removal of gender bias, statistical techniques like standard deviation and inter quartile range of the feature values are used. Details of the feature enhancement techniques may be found in [38]. The smallest class sample count is 11% of the largest class.

Please cite this article as: S. Shilaskar et al., Medical decision support system for extremely imbalanced datasets, Information Sciences (2016), http://dx.doi.org/10.1016/j.ins.2016.08.077