Expert Systems with Applications 39 (2012) 5506–5512

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

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

Random forests based monitoring of human larynx using questionnaire data

M. Bacauskiene [b], A. Verikas [a,b,*], A. Gelzinis [b], A. Vegiene [c]

[a] Intelligent Systems Laboratory, Halmstad University, Box 823, S 301 18 Halmstad, Sweden
[b] Department of Electrical & Control Equipment, Kaunas University of Technology Studentu 50, LT-51368, Kaunas, Lithuania
[c] Department of Otolaryngology, Kaunas University of Medicine Eiveniu 2, LT-50009, Kaunas, Lithuania

ARTICLE INFO

Keywords:
Random forests
Variable importance
Variable selection
Classifier
Data proximity
Human larynx

ABSTRACT

This paper is concerned with soft computing techniques-based noninvasive monitoring of human larynx using subject's questionnaire data. By applying random forests (RF), questionnaire data are categorized into a *healthy* class and several classes of *disorders* including: *cancerous*, *noncancerous*, *diffuse*, *nodular*, *paralysis*, and an overall *pathological* class. The most important questionnaire statements are determined using RF variable importance evaluations. To explore data represented by variables used by RF, the *t*-distributed stochastic neighbor embedding (*t*-SNE) and the multidimensional scaling (MDS) are applied to the RF data proximity matrix. When testing the developed tools on a set of data collected from 109 subjects, the 100% classification accuracy was obtained on unseen data in binary classification into the *healthy* and *pathological* classes. The accuracy of 80.7% was achieved when classifying the data into the *healthy*, *cancerous*, *noncancerous* classes. The *t*-SNE and MDS mapping techniques applied allow obtaining two-dimensional maps of data and facilitate data exploration aimed at identifying subjects belonging to a "risk group". It is expected that the developed tools will be of great help in preventive health care in laryngology.

1. Introduction

Multi-parameter monitoring is often seen as a way to increase the health-care quality, to reduce the feedback time, and to enable "home-care at your own responsibility". Data gathered during multi-parameter monitoring, augmented with data characterizing the specific context the subject is acting in, may be used for early detection of potential diseases. A potential disease may manifest itself as a significant deviation of a particular parameter or a particular association of parameters from the "default" expected model. Availability of cheap sensors and easiness of data transferring via wireless networks, make multi-parameter monitoring rather attractive.

Some invasive measurements are often performed to assess patient's physiological state in multi-parameter monitoring. When dealing with home-care such measurements may be too expensive and difficult to perform. On the other hand, in some cases, subject's physiological state can be assessed by efficient combination and use of all available and relevant information acquired from other non-invasive measurements. Relevant information can arise in various forms, including: numerical observations from several simple on-line sensors, off-line records containing a current and past medical history, expert opinions, environmental data, answers to statements of specifically designed questionnaires.

Laryngology is a health-care area where monitoring amounts to exploiting both invasive and non-invasive measurements. The diagnostic procedure of laryngeal diseases in clinical practice is rather complex and based on evaluation of patient's complaints, history, and data of instrumental as well as histological examination. Sequences of laryngeal images (video), voice records, electroglottogram, and questionnaire data are the main information sources used to screen for laryngeal disorders. Laryngeal images and video are probably the most important information sources when screening for laryngeal disorders (Bacauskiene, Verikas, Gelzinis, & Valincius, 2009; Verikas, Gelzinis, Valincius, Bacauskiene, & Uloza, 2007; Verikas, Uloza, Bacauskiene, Gelzinis, & Kelertas, 2009), however, the information is obtained invasively.

Patient's voice is an important source of information obtained non-invasively and used for monitoring status of patient's larynx. Automated acoustic analysis of voice is increasingly used for detecting and screening laryngeal pathologies (Gelzinis, Verikas, & Bacauskiene, 2008; Godino-Llorente & Gomez-Vilda, 2004; Hadjitodorov & Mitev, 2002; Moran, Reilly, de Chazal, & Lacy, 2006; Verikas, Gelzinis, Kovalenko, & Bacauskiene, 2010). Questionnaire data are easily obtained and also constitute an important, however under-exploited, source of information obtained non-invasively.

* Corresponding author at: Intelligent Systems Laboratory, Halmstad University, Box 823, S 301 18 Halmstad, Sweden. Tel.: +46 35 167140; fax: +46 35 216 724.
E-mail addresses: marija.bacauskiene@ktu.lt (M. Bacauskiene), antanas.verikas@hh.se (A. Verikas), adas.gelzinis@ktu.lt (A. Gelzinis), v_aurelija@yahoo.com (A. Vegiene).

0957-4174/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2011.11.070

Questionnaire data may carry information, which is not present in the acoustic or visual modalities. A questionnaire may contain a variety of statements reflecting various aspects of subject's health. Analysis of questionnaire data may be attributed to multi-parametric monitoring. Robust multi-parametric monitoring and diagnosis are, however, still rather challenging due to difficulty in interpreting hidden relations between various parameters and results of automated analysis.

This work aims at developing tools for random forests-based monitoring of human larynx using questionnaire data. Random forests is a general data mining tool and apart from prediction and data classification allows obtaining information on variable importance, proximity of different data samples, and outliers. We perform monitoring and data analysis by categorizing questionnaire data into several decision classes, namely *healthy*, *pathological*, *cancerous*, *noncancerous*, *paralysis*, *nodular*, and *diffuse*. Since questionnaire data are easy to obtain, analysis of such data can be used for preventive health care related to laryngeal diseases.

2. Data

The task considered in this work concerns automated categorization of questionnaire data into several decision classes. Several categorization alternatives are considered: (i) *healthy* and *pathological*; (ii) *healthy*, *paralysis*, *diffuse*, and *nodular*; (iii) *healthy*, *cancerous*, and *noncancerous*. *Diffuse* and *nodular* are two groups of mass lesions of vocal folds. Mass lesions of vocal folds could be categorized into six subclasses namely, *polypus*, *papillomata*, *carcinoma*, *cysts*, *keratosis*, and *nodules*. This categorization is based on clinical signs and histological structure of the mass lesions of vocal folds. Mass lesions of vocal folds can be *cancerous* lesions—*carcinoma* and *noncancerous* lesions—*nodules*, *polyps*, *cysts*, *papillomata*, and *keratosis*. The diagnosis in each case was confirmed by histological examination of laryngeal specimens removed during endolaryngeal microsurgical intervention.

In total, questionnaires were collected from 109 subjects. There are 26 questionnaires in the *healthy* class, 83 in *pathological*, 29 in *diffuse*, 46 in *nodular*, 8 in *paralysis*, 29 in *cancerous*, and 54 in the *noncancerous* class.

There are 47 questionnaire statements in total. Below given is a list of the most important questionnaire statements selected using the developed algorithms.

2.1. The most important questionnaire statements

The following questionnaire statements (components x_i of the data vector $\mathbf{x}$) were selected as being the most important ones for data classification into the aforementioned classes:

2. Levels of vocal usage (four levels);
3. average duration of speech use (hours/day);
4. average duration of speech use (days/week);
6. smoking (yes/no);
7. smoking duration (years);
8. smoked cigarets/day;
11. subjective hoarseness assessment by the patient on a scale ranging from 0 (“no hoarseness”) to 100 (“severe hoarseness”);
12. subjective voice assessment by the patient on the Visual Analogue Scale ranging from 0 to 100;
13. voice handicap progressing (four grades);
17. subjective assessment of daily experienced stress level on a scale ranging from 0 (“no stress”) to 100 (“very much stress”);
22. frequency of singing (five grades);
26. frequency of talking/singing in a smoke-filled room (five grades);
35. subjective assessment of experienced discomfort due to voice disorder on a scale ranging from 0 (“no discomfort”) to 100 (“huge discomfort”);
38. subjective assessment of “too weak voice” on a scale ranging from 0 (“no”) to 100 (“very clear”);
41. subjective assessment of repetitive “loss of voice” on a scale ranging from 0 (“no”) to 100 (“very clear”);
42. subjective assessment of reduced voice range on a scale ranging from 0 (“no”) to 100 (“very distinctly”);
43. subjective assessment of reduced ability to sing on a scale ranging from 0 (“no”) to 100 (“very distinctly”);
44. frequency of voice cracks or aberrant voice on a scale ranging from 0 (“no”) to 100 (“very often”).

3. Weak learners and random forests

Let us assume that given is a set of training data $\mathcal{X}_t = \{(\mathbf{x}_m, y_m), m = 1, \ldots, M\}$, where $\mathbf{x}_m$ is an input observation and y_m is a predictor output. A weak learner can be created using the training set $\mathcal{X}_t$. A weak learner is a predictor $f(\mathbf{x}, \mathcal{X}_t)$ having low bias and high variance. By randomly sampling from the set $\mathcal{X}_t$, a collection of weak learners $f(\mathbf{x}, \mathcal{X}_t, \boldsymbol{\theta}_k)$ can be created, where $f(\mathbf{x}, \mathcal{X}_t, \boldsymbol{\theta}_k)$ is the kth weak learner and $\boldsymbol{\theta}_k$ is the random vector selecting data points for the kth weak learner. By applying bootstrap sampling, for example, to generate $\boldsymbol{\theta}_k$, about two-thirds of the data points are used by each weak learner. About one-third of the observations are out of the bootstrap sample or out-of-bag (OOB). The $\boldsymbol{\theta}_k$ are independent and identically-distributed, i.i.d.

It can be shown that combining i.i.d. randomized weak learners into a committee by averaging, leaves the bias approximately unchanged while reduces the variance by a factor of $\bar{\rho}$ – the mean value of the correlation between the weak learners (Breiman, 2001). Thus, if correlation and bias of i.i.d. randomized weak learners are kept low, big reduction in test set error can be obtained.

Random forest (RF) is a committee of weak learners for solving prediction (both classification and regression) problems. In RF, a decision tree, i.e. CART (classification and regression trees), is used as a weak learner. When solving classification problems, RF prediction is the un-weighted majority of class votes. As the number of trees in RF increases, the test set error rates converge to a limit, meaning that there is no over-fitting in large RFs (Breiman, 2001). Low bias and low correlation are essential for accuracy. To get low bias, trees are grown to maximum depth. To achieve low correlation, randomization is applied:

i. Each tree of RF is grown on a bootstrap sample of the training set.
ii. When growing a tree, at each node, n variables are randomly selected out of the N available.
iii. Usually, $n \ll N$. It is suggested starting with $n = \lfloor \log_2(N) + 1 \rfloor$ or $n = \sqrt{N}$ and then decreasing and increasing n until the minimum error for the OOB data set is obtained. At each node, only one variable, providing the best split, is used out of the n selected.

In RF, n is the only parameter to be selected experimentally. RF can handle thousands of variables of different types with many missing values. For a tree grown on bootstrap data, the OOB data can be used as a test set for that tree. As the number of trees increases, RF provides an OOB data-based unbiased estimate of the test set error. OOB data are also used to estimate importance of variables. These two estimates (test set error estimate and variable importance) are very useful byproducts of RF. Random forests have been successfully applied in a variety of fields (Coussement & Van