King Saud University

Applied Computing and Informatics

www.ksu.edu.sa
www.sciencedirect.com

ORIGINAL ARTICLE (INVITED)

On classification in the case of a medical data set with a complicated distribution

Martti Juhola [a,*], **Henry Joutsijoki** [a], **Heikki Aalto** [b], **Timo P. Hirvonen** [b]

[a] *Computer Science, School of Information Sciences, University of Tampere, Tampere 33014, Finland*
[b] *Department of Otorhinolaryngology & Head and Neck Surgery, University of Helsinki and Helsinki University Central Hospital, Helsinki 00029 HUS, Finland*

Received 27 February 2014; accepted 14 March 2014
Available online 25 March 2014

KEYWORDS

Data mining;
Data cleaning;
Complicated data distributions;
Classification

Abstract In one of our earlier studies we noticed how straightforward cleaning of our medical data set impaired its classification results considerably with some machine learning methods, but not all of them, unexpectedly and against intuition compared to the original situation without any data cleaning. After a more precise exploration of the data, we found that the reason was the complicated variable distribution of the data although there were only two classes in it. In addition to a straightforward data cleaning method, we used an efficient way called neighbourhood cleaning that solved the problem and improved our classification accuracies 5–10%, at their best, up to 95% of all test cases. This shows how important it is first very carefully to study distributions of data sets to be classified and use different cleaning techniques in order to obtain best classification results.

[*] Corresponding author. Tel.: +358 40 1901716; fax: +358 3 2191001.
E-mail address: Martti.Juhola@sis.uta.fi (M. Juhola).

Peer review under responsibility of King Saud University.

ELSEVIER Production and hosting by Elsevier

2210-8327 © 2014 King Saud University. Production and hosting by Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.aci.2014.03.001

1. Introduction

In our earlier research we developed a signal analysis method for nystagmic eye movements investigated in otoneurological tests (Juhola et al., 2009, 2011). For the automatic analysis of such signals poor or invalid nystagmic eye movements should correctly be separated from valid nystagmic eye movements, because valid eye movements can only be used for the data analysis needed for the diagnostics of otoneurological patients. Typically, invalid nystagmic eye movements are corrupted by noise or artefacts. Thus, we have also studied the classification of nystagmic eye movement candidates into invalid and valid, hereafter called the rejected and accepted, on the basis of machine learning methods (Juhola et al., 2013). We then observed how their complicated distribution made the classification task difficult and attempted to reduce the greater subset (class) of the rejected eye movement candidates in a learning set, which was performed by cleaning away a part from them. Surprisingly, a simple cleaning process impaired classification results of some of the machine learning methods applied. We realized that the reason for such a seemingly conflicting situation originated from the complicated variable distribution of the data (Juhola et al., 2013).

In order to define which distribution of two classes is seen as simple or complicated we refer to Fig. 1. A simple distribution is where the centre (computed as means for all variables) of each class is located inside its own area including most elements of that class. This is described in Fig. 1(a). A complicated distribution is depicted in Fig. 1(b), where the centre of one class is outside its own area. Such complicatedness could be defined in various ways, but it is essential that we cannot then base data cleaning on distances from the class centres.

Nystagmus is formed from repeated, reflexive eye movements occurring as to-and-for beats that can be measured in the horizontal, vertical and torsional directions with two eye movement video cameras, one for each eye. A hypothetic nystagmic eye movement beat is seen in Fig. 2 and an actual signal of several nystagmic beats in Fig. 3. A nystagmic beat includes the slow phase immediately followed by the fast phase in order to return the eye in the opposite direction. Nystagmic beats are repetitive and their configurations vary even in the course of short measurement times. A healthy subject performs nystagmic eye movements, for example, when he or she is sitting in a moving train and looks at changing (relatively close) views through a window. This is called optokinetic nystagmus because of the stimulation. Caloric nystagmus is induced by injecting a small quantity of cool or warm (37 ± 7 °C) air or water into the ear canal of a subject. In regard to some otoneurological disorders or diseases, head shaking or head movement can provoke nystagmus and even spontaneous nystagmus may appear in vestibular patients. Congenital nystagmus also exists (Hertle and Dell'Osso, 1999). The slow phase features (variables) of nystagmus are important for the diagnostics of vestibular neuritis, positional vertigo, vestibular schwannoma and Menière's disease. The fast phase features of nystagmus are important for