Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 98 (2016) 467 – 472

Procedia Computer Science

International Workshop on Data Mining on IoT Systems (DaMIS16)

Applying Mining Techniques to Analyze Vestibular Data

Domenico Mirarchi*[a], Claudio Petrolo[c], Giovanni Canino[a], Patrizia Vizza[a], Salvatore Cuomo[b], Giuseppe Chiarella[a,c], Pierangelo Veltri[a]

[a] *University of Catanzaro - Department of Medical and Surgical Sciences,viale Europa, Catanzaro 88100, Italy*
[b] *University of Naples Federico II - Department of Mathematics and Applications,via Cintia, Napoli 80126,Italy*
[c] *University of Catanzaro - U.O. Audiology and Phoniatrics,viale Europa, Catanzaro 88100, Italy*

Abstract

The vestibular apparatus allows to perform audiological and equilibrium human functions and to capture movements with respect to gravity. Damages to the vestibular system causes diseases that can be measured by using Vestibular Evoked Myogenic Potentials (VEMPs) test. The test produces a lot of data that has to be collected and analyzed to allow a disease study and classification. We propose a framework that includes algorithms able to perform pathology distribution and classification. It has been tested on electronic patient records loaded from the University Hospital database. The software allows to manage the structure and framework and a blind application of one of the available classification techniques shows a relation among gender and vestibular apparatus disease.

Peer-review under responsibility of the Program Chairs

Keywords: vestibular disease; data mining; disease classification

1. Introduction

The vestibular human apparatus is the inner part of the ear (also called labyrinth), constituted by bones and soft tissues. It is the sensory system that allows to detect the position and motion of the head, i.e., rotation and motion, allowing balance and spatial orientation. While head is moving, the vestibular apparatus stabilizes the eyes and adjusts neck and body muscle tone during movements [15] to send information useful to balance movements. The brain uses such information to analyse movements and to guide muscles to dynamically balance body cinematic.

Studying vestibular functions is necessary to identify and characterize vestibular disorders that may be responsible of human pathologies related for instance to equilibrium, head tilt, asymmetrical ataxia, or nystagmus [16]. Specific symptoms of vestibular disorders can be identified by studying results by measuring Vestibular Evoked Myogenic Potentials (VEMPs). The VEMPs is a test performed with external electrodes to evaluate the Vestibular system [18],

* Corresponding author. Tel.: +39-0961-3694149; fax: +39-0961-369-4073.
E-mail address: d.mirarchi@unicz.it; canino@unicz.it; salvatore.cuomo@unina.it; claudiop26@virgilio.it; vizzap@unicz.it; chiarella@unicz.it; veltri@unicz.it

Peer-review under responsibility of the Program Chairs
doi:10.1016/j.procs.2016.09.076

that can be executed in cervical and ocular zone to measure the related myogenic potentials, called respectively cervical VEMP (cVEMPs) and ocular VEMP (oVEMPs) potentials. This test reveals saccular function in response to airconducted sound stimulation using surface electrodes over the sternocleidomastoid muscles; furthermore, it reveals vestibular function of inferior nerve, and vestibulocollic connections[1]. Testing the vestibular system requires dierent analysis, part of them conducted by physicians through a patient visit, others by using medical devices and instruments. Each test produces a lot of data and it can be repeated several times, and there is no standard model for data integration and storing. Nevertheless, to the best of our knowledge there is no general purpose software able to collect heterogeneous data into a single information data structure able to analyse information coming from different sources (both different devices and physician notes and observations) in a unique platform.

The here proposed framework aims to collect and analyse data coming from vestibular system to obtain additional information and to support physicians while dening diagnosis. Therefore, algorithms (such as data mining, machine learning based ones, decision support systems) can be useful to support correct diagnosis of specific pathology related to a patient[4,3]; in particular, the data can be used to build classification models useful for either diagnosis, prognosis or treatment planning. Moreover, data mining algorithms can be used for predictive purpose to find interesting patterns in the data, as well as clusters and subgroups of data. Known techniques adopted for healthcare are: (i) neural networks, (ii) decision trees, (iii) genetic algorithms and (iv) nearest neighbor method. Specically, articial neural networks use a learning process similar to the human brain: each connection inside the neural networks becomes important and all connections together realize the input-output system[8]. A decision tree is used for data classification and it is a graphical representation of the relations that exist between the data in the database; this technique is mainly used in classication and prediction topics, and it is a simple way for representing the knowledge[9]. Genetic algorithms are based on the principle of genetic modification and are inspired by the principles observed in natural evolution[8,14,7]. Lastly, nearest neighbour method is used for data classification; it analyzes all the data in the database to find a subset of instances that are the best fit and uses this subset to predict the outcome. This method is used, for example, to detect efficiency in the diagnosis of heart diseases[17]. Bayesian based methods have been used for health data analysis. For example, author in[13] studies the simply application for health surveillance data. Authors in[6] use Bayesian networks for representing statistical dependencies for gene data. In[12] author adopts a method to model the uncertainties and to allow integration between biomedical and clinical background knowledge. Also, parallel analysis and collaborations among different laboratories can be performed similarly to[2].

In this paper we dene a software system to acquire and integrate data coming from devices able to measure vestibular related disease, and to apply data mining algorithms to data coming from VEMPs test. Anonymized data and clinical notes gathered from the audiological medicine unit of the University of Catanzaro Magna Graecia (UMG) have been used to populate the system; 1976 patients have been enrolled to test and populate the system. We here present results in applying bayesian methods on 400 patients of enrolled 1976 patients to predict a disease based on gender information: preliminary results show that probability of finding disease is higher in female than male. The results are not deep in terms of parameters, due to the fact that available data in the patient records used in database are poor in terms of information. Thus it has not been possible to perform deep tests for lacking of information. Nevertheless, the system is now hosting new patient records and it is in use to the audiological unit; the system can be considered as a valid support to define a knowledge base information representing the first necessary step for further analysis.

2. Architecture

The system must allow the definition of patient information as well as the acquisition of data coming from medical devices and the collection of information from vestibular tests. It must fulfill the following requirements:

- it must be developed on the basis of an architectural model that defines an operating infrastructure to support interoperability solutions between the medical practitioner and the software component;
- it must identify information about the patient, standardized infrastructure and technological aspects, and provide data security;
- it must be a valid support for vestibular diseases, providing a platform for storing and management of health data.

The system is used for the insertion of patient data, amnesiac records, tests results, annotation of diagnosis, therapy and different disorders associated with results coming from external tests.

The Entity Relationship Diagram (ERD) has been defined by using clinical information. ERD has been realized with MyAQL Workbench software; this software is a unified visual development and administration platform that includes advanced tools for database modeling and design, query development and testing, server configuration and monitoring, user and security administration, backup and recovery automation, and audit data inspection. The ERD allows to define the database containing the following tables: patient, examination, treatment, diagnosis and general information, device data and information. Figure 1 represents some information regarding tables and an extracted set of data instance.

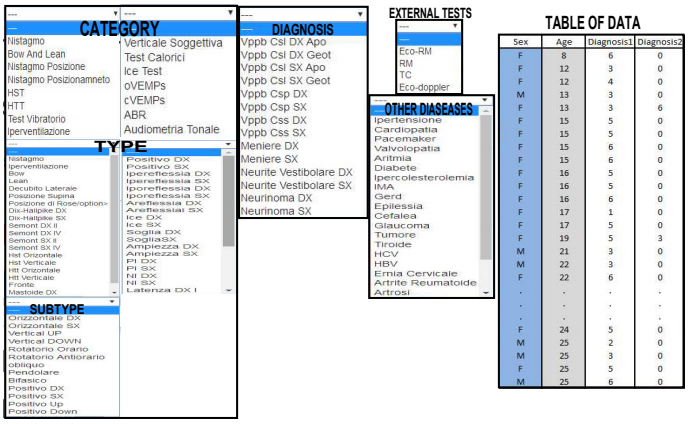

Fig. 1. Views and data tables

The database instance is performed on a PostGreSQL instance, with spatial extension. The system has been developed by using a web application with components and graphical user interface that are represented in Figure 2.

Fig. 2. Login screen and Main GUI

The main page reports three operation on the patient data that can be performed by physician, such as search of data patient stored in database, insertion of a new patient and extraction of a patient list filtered by specific pathologies. In particular, the search of patient data is used to acquire and verify the information regarding patients who have already made a visit and a VEMPs test. Using the list of patient the operator can view the date of the examination acquisition and the name of the patient.

The insertion of patient data allows to acquire data coming from external data sources and clinical devices. This phase is divided in more steps, as reported in Figure3.

The first GUI shows the insertion of personal data relating to the patient; in particular, the personal data are fitted with related residence information of the patient. The second GUI is used for the insertion of data relating the medical tests; these tests have been divided into primary and secondary tests. For any test, the physician can choose

category, type and sub type; the categories are used to identify the general name of the test, specifying name and output of the test. Diagnosis and therapy are inserted in the third step; diagnosis is selected from a possible number of solutions given by the doctor and, in particular, there are three possible areas of selection. The last GUI is used for the insertion of external associated tests and related pathologies of the patient; the physicians can select a set of associated pathologists for the patient. The number of selectable diseases can be increased to improve the set for a better choice selection.

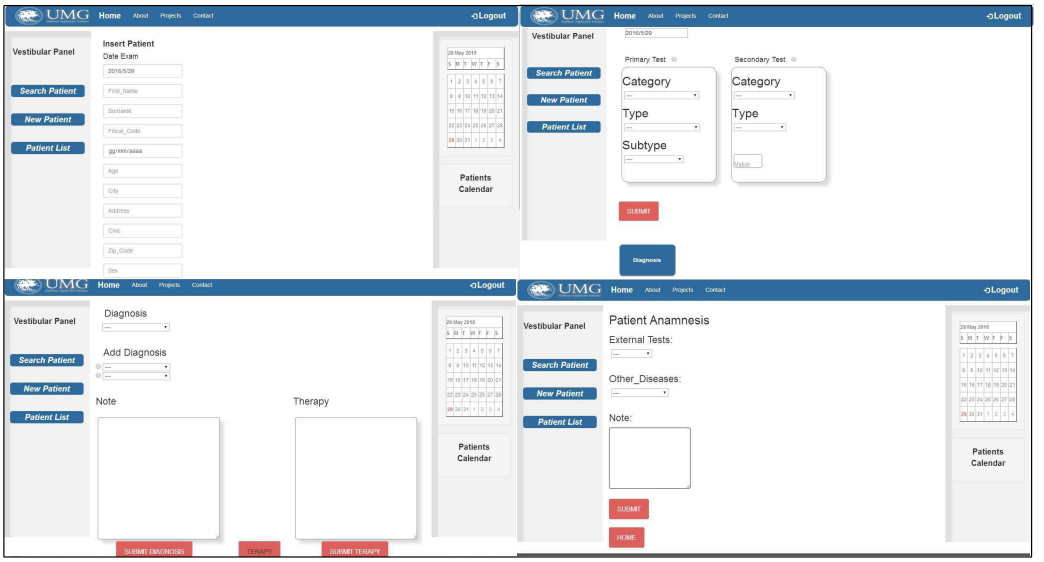

Fig. 3. Acquiring data from external sources and devices

3. Analysis module

The system includes a section to perform analysis based on the selection of data mining based algorithm. We include in the first version a bayesian based model. The Bayesian based methods can be applied to any data and they provide most flexibility for data analytic models[5]. These methods have the follow advantages: (i) there are no p values in Bayesian analysis, (ii) inferences provide rich and complete information regarding all parameters, and (iii) models can be customized for different types of data. Bayesian methods are used also to estimate the probability in any kind of experiment[10]. The data have been organized in a file used as input for the classifier; the data have been analyzed to have a comparison metric with the results coming from the model. The methods are implemented in supervised induction computational tasks, where the performance target is to accurately predict the class of test instances and in which the training instances involve to class information.

The selected one is fairly intuitive and is based on the minimization of the following cost function:

$$CM(x_1, x_2, \cdots, x_n) =$$
$$= \arg\max_z p(Z = z) \prod_{i=1}^{N} p(X_i = x_i | Z = z)$$

where Z is a dependent class variable and $X_1 \cdots X_n$ are several feature variables.

The classifier is based on the computation of individual conditional probabilities for each values of the class variable Z and for each feature $p(X_i|Z_j)$. The class, given by Bayesian classifier, is the one for which we have the largest product of the probabilities. The Maximum Likelihood Estimation Method[11] is used to determine the individual conditional probabilities.

Based on that, we have elaborated data regarding 1976 patients stored in the database to extract useful information to identify the relation between patients and diagnosis, specifically between the gender of patients and diagnosis. The goal is to verify if the bayesian method can be used as possible predictor for vestibular system disease. In particular, the experiment consists in the analysis of the data stored to extract a probability that a vestibular pathology is more present in male or female patients. In database, data patient have been considered and divided in the following categories: (i) patient data, (ii) data derived by VEMPs exams, (iii) data coming from patient information, and (iv) data produced by medical consideration. The tables have been structured in columns that correspond to the analyzed attributes: (i) sex, (ii) age, (iii) first diagnosis and (iv) second diagnosis. The study has produced results in term of predisposition and propensity of women patients to contract the diseases of the auditory system. The system includes libraries that invoke the R software for statistical computing and graphics. In Figure 4, the phase of calculation for running the above reported formula for bayesian analysis is reported. Results are compared with applying of Weka module, an open source software for automatic learning. In Figure 5 the verification phase of the classification is reported. By comparing the output produced by the two software, we verify that the results are matching showing a predisposition and a propensity of women to lodge diseases of the auditory system.

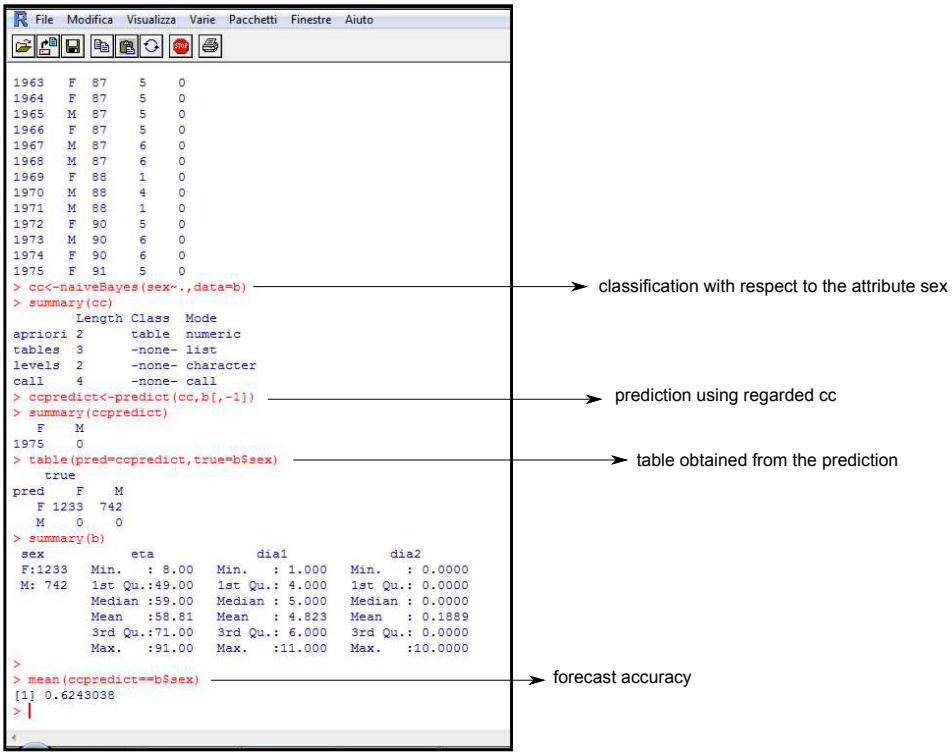

Fig. 4. Evaluating statistics module

4. Conclusions

The proposed contribution represents a first analysis to extract relevant and useful information of vestibular pathology regard to the sex. New analysis performed by alternative methods could improve the quality and the quantity of data to extract more information about the development of vestibular pathology. The system has been tested by crawling and loading data extracted from anonymized patient records. Most of the clinical records have been reported in paper format thus lacking of data and medical devices data. Nevertheless the system is now used by the audiological unit and it stores whole information from each patient analysis. New and more complete data can be stored in the

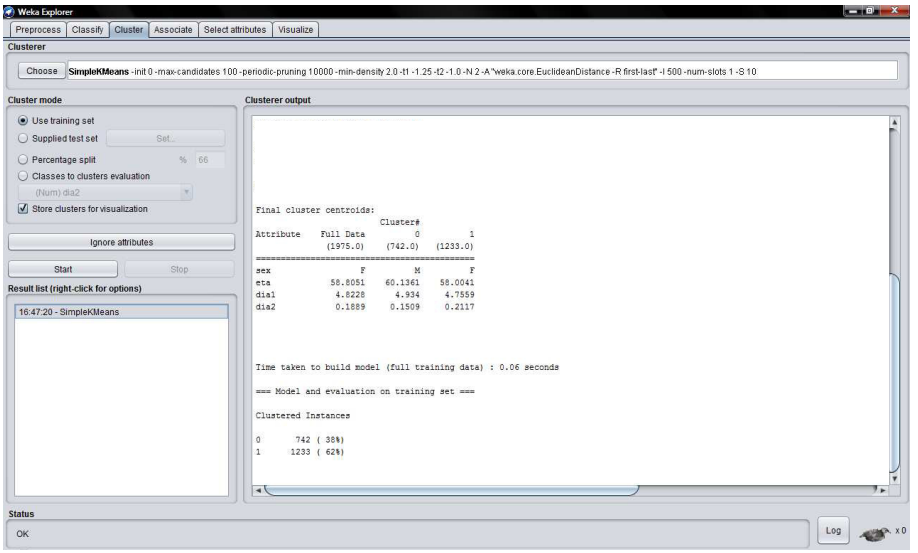

Fig. 5. Weka verification phase

system so that different mining module can be used to give additional information of different nature to better understand the mechanisms of disease diffusion. Moreover, a geographic module is just implemented into the database but its functionalities can be included to monitor the diffusion of vestibular pathologies in geographic areas.

References

1. K. Brantberg. Vestibular evoked myogenic potentials (vemps): usefulness in clinical neurotology. In *Seminars in neurology*, volume 29, pages 541–547, 2009.
2. M. Cannataro, P. H. Guzzi, and P. Veltri. Impreco: Distributed prediction of protein complexes. *Future Generation Computer Systems*, 26(3):434–440, 2010.
3. M. Cannataro, P. H. Guzzi, and P. Veltri. Using ontologies for querying and analysing protein-protein interaction data. *Procedia Computer Science*, 1(1):997 – 1004, 2010.
4. K. J. Cios and G. W. Moore. Uniqueness of medical data mining. *Artificial intelligence in medicine*, 26(1):1–24, 2002.
5. C. Erdman, J. W. Emerson, et al. bcp: an r package for performing a bayesian analysis of change point problems. *Journal of Statistical Software*, 23(3):1–13, 2007.
6. N. Friedman, M. Linial, I. Nachman, and D. Pe'er. Using bayesian networks to analyze expression data. *Journal of computational biology*, 7(3-4):601–620, 2000.
7. S. Greco, C. Molinaro, and I. Trubitsyna. Logic programming with function symbols: Checking termination of bottom-up evaluation through program adornments. *Theory and Practice of Logic Programming*, 13(4-5):737–752, 2013.
8. S. Gupta, D. Kumar, and A. Sharma. Data mining classification techniques applied for breast cancer diagnosis and prognosis. *Indian Journal of Computer Science and Engineering (IJCSE)*, 2(2):188–195, 2011.
9. F. S. Khan, R. M. Anwer, O. Torgersson, and G. Falkman. Data mining in oral medicine using decision trees. *World Academy of Science, Engineering and Technology*, 37:225–230, 2008.
10. J. K. Kruschke. What to believe: Bayesian methods for data analysis. *Trends in cognitive sciences*, 14(7):293–300, 2010.
11. R. J. Little and D. B. Rubin. The analysis of social science data with missing values. *Sociological Methods & Research*, 18(2-3):292–326, 1989.
12. P. Lucas. Bayesian analysis, pattern analysis, and data mining in health care. *Current opinion in critical care*, 10(5):399–403, 2004.
13. D. Madigan. Bayesian data mining for health surveillance. *Spatial and syndromic surveillance for public health*, pages 203–221, 2005.
14. P. S. Ngan, M. L. Wong, W. Lam, K. S. Leung, and J. C. Cheng. Medical data mining using evolutionary computation. *Artificial Intelligence in Medicine*, 16(1):73–96, 1999.
15. D. E. Parker. The vestibular apparatus. *Scientific American*, 1980.
16. K. L. Schunk. Disorders of the vestibular system. *Veterinary Clinics of North America: Small Animal Practice*, 18(3):641–665, 1998.
17. M. Shouman, T. Turner, and R. Stocker. Applying k-nearest neighbour in diagnosing heart disease patients. *International Journal of Information and Education Technology*, 2(3):220, 2012.
18. M. S. Welgampola and J. G. Colebatch. Characteristics and clinical applications of vestibular-evoked myogenic potentials. *Neurology*, 64(10):1682–1688, 2005.