33rd Annual International Conference of the IEEE EMBS
Boston, Massachusetts USA, August 30 - September 3, 2011

Heterogeneous Data Fusion and Intelligent Techniques Embedded in a Mobile Application for Real-Time Chronic Disease Management

Christos Bellos, Athanassios Papadopoulos, Roberto Rosso, and Dimitrios I. Fotiadis, *Senior Member, IEEE*

Abstract—**CHRONIOUS system is an integrated platform aiming at the management of chronic disease patients. One of the most important components of the system is a Decision Support System (DSS) that has been developed in a Smart Device (SD). This component decides on patient's current health status by combining several data, which are acquired either by wearable sensors or manually inputted by the patient or retrieved from the specific database. In case no abnormal situation has been tracked, the DSS takes no action and remains deactivated until next abnormal situation pack of data are being acquired or next scheduled data being transmitted. The DSS that has been implemented is an integrated classification system with two parallel classifiers, combining an expert system (rule-based system) and a supervised classifier, such as Support Vector Machines (SVM), Random Forests, artificial Neural Networks (aNN like the Multi-Layer Perceptron), Decision Trees and Naïve Bayes. The above categorized system is useful for providing critical information about the health status of the patient.**

I. Introduction

THE necessity of the system rises from the need to perform medical procedures and monitor chronically ill people remotely in the most efficient way according to the most updated international guidelines [1], [2]. Clinicians need to be informed about the patient profile related information (history, latest measurements, prescriptions, appointments) in a frequent base and prescribe a detailed patient plan (e.g. diet, drugs, exercises, lab tests and questionnaires). Additionally, clinicians should be able to set the patients' alarms and make queries to the most updated medical knowledge in a regular protocol.

In order to fulfill the aforementioned necessity that was elicited from the user requirements procedure, CHRONIOUS platform captures various types of signals in order to make in real-time a preliminary assessment of the patient's health status [3]. Heterogeneous data are being gathered from different sources, like the wearable sensors platform, SD interfaces and SD database. These data are being gathered in the Intelligent Core of the integrated SD application and forwarded into four independent components, feature extraction, patient inputted data analyzer, decision support system and severity estimation [4]. The wearable sensors platform contain a 3-lead Electrocardiogram (ECG), a microphone as a context-audio sensor, a pulse oximeter, two respiration bands (thorax and abdominal), an accelerometer and a sensor for measuring humidity as well as body and ambient temperature. The signals acquired from the wearable sensors are being transferred via Bluetooth to the SD and analyzed by the Feature Extraction component. In case Abnormal Situation Tracking is activated, placed at the Data Handler, it recognizes an abnormal condition (by containing thresholds and applying simple rules) and thus a signal is being sent to SD and restores its functions to normal mode. Then, the Feature Extraction module is being triggered and incoming signals are being processed. Alternatively, an internal clock can activate Feature Extraction component in order to perform scheduled signal analysis and thus the extracted attributes are being forwarded to the decision support system and the severity estimation procedure.

The patient inputted data analyzer contains three independent processes:

1) Food Intake Analysis for the calculation of the total ingredients and calories consumed for each food intake depending on the user selections (i.e. Calcium, Sodium, Potassium, Water, Carbohydrates, Total Lipids, etc.).

2) Drug Intake Analysis for the analysis of the drug intake acknowledgements (on-time, delayed or missed intakes) to produce an overall assessment of patient's adherence to clinician's prescription.

3) Questionnaire Analysis for the analysis of patient's estimation to his/her health status.

Patient inputted data and extracted features from signals acquired by the wearable sensors are entering the heterogeneous data fusion analyser in order to be combined and annotated in the time domain. Before the classifier testing phase being executed, a preliminary phase of feature selection and sensitivity analysis is being triggered. Several feature selection algorithms have been applied and will be further analyzed in the respective section.

DSS is embedded at the SD integrated application. This component decides on patient's current health status by combining the available heterogeneous information. In case no abnormal situation has been tracked, the DSS takes no action and remains deactivated until next pack of data are being acquired.

The DSS that has been implemented is an integrated

Manuscript received April 15, 2011. This work has been partly funded by the European Commission through IST Project FP7-ICT-2007–1–216461 www.chronious.eu.

C. Bellos, A. Papadopoulos, D.I. Fotiadis are with FORTH BRI Foundation for Research and Technology - Hellas, Biomedical Research, Ioannina Greece (e-mail: cbellos@cc.uoi.gr, thpapado@cc.uoi.gr, fotiadis@cs.uoi.gr). R. Rosso is with TESAN Telematic & Biomedical Services S.p.A., Vicenza Italy (e-mail: rosso@tesan.it).

978-1-4244-4122-8/11/$26.00 ©2011 IEEE

classification system with two parallel classifiers, combining an expert system (rule-based system) and a supervised classifier, where several tests have been performed using various classifiers like Support Vector Machines (SVM) [6], Decision Trees (Random Forest [7], [16], C4.5 Decision Tree [15], PARTial Decision Tree [13], [14], artificial Neural Network (Multi-Layer Perceptron) [8], and Naïve Bayes [11].

II. Materials and Methods

The Feature Extraction and Heterogeneous Data Fusion as well as the DSS form the core of the application logic of the PDA system.

A. Feature Extraction and Heterogeneous Data Fusion

TABLE I
EXTRACTED FEATURES AND THEIR SOURCE

Wearable Sensor	Features Extracted
ECG Sensor	Mean QR Distance + deviation Mean RS distance + deviation Mean RR + deviation Mean HR + deviation LF/HF
Respiration Sensor	Resp Frequency and amplitute Inhalation-Exhalation Duration
Acceleration Sensor	Min of Standing, lying Num of detected Steps and falls
Audio Sensor	Events and dBs of Cough Events of Snoring Environmental Noise (dBs)
Other Wearable Sensors	Body and ambient Temperature Environmental Humidity SpO2 (Pulse Oximeter)
Food Intake	Total Calories, Lipids Carbohydrates (CHO) Other (e.g. Calcium, Sodium)
Medication Intake	Characterization of adherence at clinician's prescription (drug intake on time, drug received later than scheduled, missed drug)
Activity Data	Activity Energy Expenditure (Calories) Activity additional comments (e.g. Feel sick, nausea, muscle pain, etc.)
Questionnaires (disease scpecific questions)	4 mental questions 9 COPD Questions (for COPD patients) and 6 CKD questions (for CKD patients)
External Sensors	Breathing Asynchrony (Spyrometer) Breathing Rapidly (Spyrometer) Systolic-Diastolic Blood Pressure Blood Glucose Weight (Body Weight Device)
Environmental Sensors	Increased air particles Presence of smoke
Database (ClinicalData)	Smoking Status Exposure to smoke and dust (Yes, No) Dyspnea – chronic cough and sputum
Database (PatientInfo)	Patient Information (e.g. age, gender)

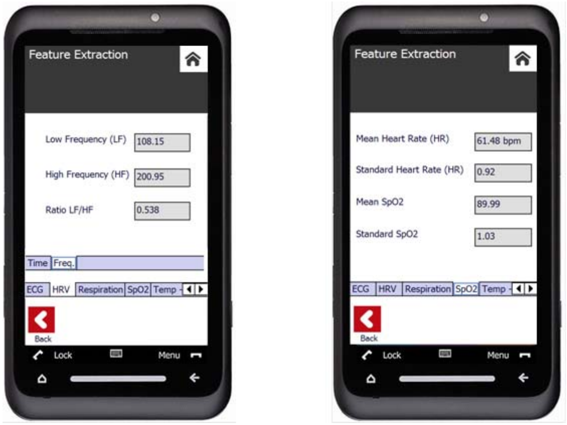
As has already been mentioned in the Introduction section, the input of the DSS is a multi-dimensional vector of data, formed from data acquired by three main sources:

1) Wearable sensors: Sensors that are integrated to the wearable jacket record signals in both discrete and continuous format.

2) Database: queries to the database return data (e.g. lifestyle stored information or demographics data) either to a DataTable or to a string variable.

3) SD Graphical User Interface: Patient enters information using the interfaces of the SD regarding food intake, medication, activity events and responses to disease or/and mental specific questionnaires.

These data are being analyzed and features vectors are being extracted either by the Feature Extraction component or by the patient inputted data analyzer. The extracted features are presented in Table I. After being extracted, the features enter to the heterogeneous data fusion component which fuses all available information and forms the multi-dimensional vector to feed the DSS and trigger the developed classifiers. A sample of the extracted features is being displayed where the attributes are being presented into the interface of the Smart Device.



Low Frequency (LF) 108.15
High Frequency (HF) 200.95
Ratio LF/HF 0.538

Mean Heart Rate (HR) 61.48 bpm
Standard Heart Rate (HR) 0.92
Mean SpO2 89.99
Standard SpO2 1.03

Fig. 1. Heart rate variability related features and SpO2 extracted

Before the classifiers are being executed a preliminary phase of feature selection is being triggered. Several attribute evaluator algorithms have been applied and will be further analyzed in the respective section [5]. The most important approaches that generate the best results for our datasets are the following [17]:

1) Principal Component Analysis (PCA): This algorithm is being used in conjunction with a Ranker search.

2) GainRatioAttributeEval: This algorithm measures the gain ratio and evaluates the worth of each attribute.

3) Correlation-based Feature Subset Selection: Considers the individual predictive ability of each feature and evaluates the worth of a subset of attributes.

4) ConsistencySubsetEval: Evaluates the worth of a subset of attributes.

5) WrapperSubsetEval: Evaluates the attribute sets by using a learning scheme, where Naïve Bayes is being used as the required classifier.

B. Decision Support System

The aim of the DSS [9], [12] component is twofold; initially constructs the training model, trains the algorithms and afterwards classifies the health episodes according to five different levels of severity on an application embedded in the SD. This twofold aim is being accomplished through three main phases that have been developed.

The training phase of the algorithms has been developed in C# using .NET 3.5 SP1 Framework and deployed under the Visual Studio 2010 Environment. The application is divided into two independent projects; feature extraction and training models. The Feature Extraction project contains a form in order to load the data and visualize the extraction of the features from sensors' acquired data that have already been stored in the Central Database. After feature extraction, the algorithms are being trained and the results are being stored to simple text files in order to be downloaded to the SD.

The update phase guarantees the efficiency of the DSS and enhances the accuracy and the personalization of outcomes. The updating phase is being triggered by a timer, which has been developed in order to trigger an UPDATE query to the SQL Server Database.

The testing phase that is embedded in the SD CHRONIOUS application has been developed in .NET 3.5 Compact Edition SP1 and under the Visual Studio 2008 Development Environment. When the SD application is initiated, recent data are being acquired from the Database using the SQL Outer Join Query in order to obtain data that have been stored into various fields of different tables. Two classes have been developed in order to integrate and fuse the heterogeneous data that have been acquired from the database. The DataIntegration() class splits the dataset to five different time zones in order to annotate data in the time domain and store to one line of the generated vector only data that appear in the same time zone. On the other hand, the DataFusion() class initially forms the input to the Feature Extraction component and then fuses data to an integrated vector.

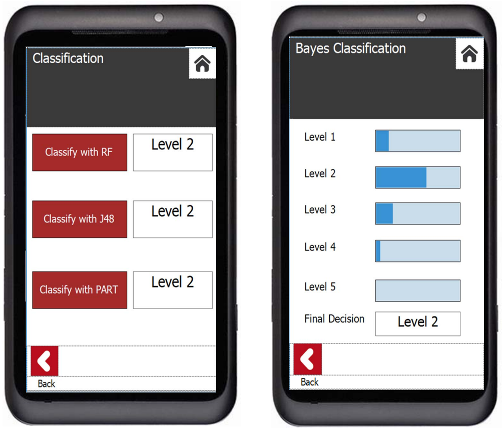


Fig. 2. Snapshots from the Smart Device where the result of four classifiers (Random Forest, J48 which is the C4.5, PART and naïve bayes) is being presented. The Final decision to all classifiers is that the health status of the patient is identified as "level 2".

After extracting required features and fusing heterogeneous data, the generated integrated input vector is feeding the various classifiers that have been developed in order to assess the severity of the health status. The Fig. 2 displays a sample of the result of the classification for a Level 2 patient's health status.

III. PROTOCOL DESIGN

In our preliminary analysis we have formatted and used three datasets with different features from Chronic Obstructive Pulmonary Disease (COPD) and Chronic Kidney Disease (CKD) patients.

The first dataset, named QuestData_COPD, contain 58 features that have been acquired through responses to questionnaires both from Clinicians and 20 COPD patients. The features are acquired through one day of recording while clinicians provided also an extra feature with specifying the annotated level of a patient for the specific measurement. Also, the dataset is split into training and testing dataset. The 10 folds cross-validation is applied as well as the percentage split (we have applied 66% for splitting the data set: 66% of the data form the training set and 33% of the data form the test set).

Finally, the classification is being performed by applying various aforementioned algorithms and the results are being presented in the next table, in order to present a comparison between the methods according to the percentage of the correctly classified instances.

The second dataset, named QuestData_CKD, contain 35 features that have been acquired through responses to questionnaires both from Clinicians and 16 CKD patients. The features are acquired through five days of recording while clinicians provided also an extra feature with specifying the annotated level of a patient for the specific measurement.

The third dataset, named FileData, contain 23 features that have been acquired through de-noising and feature extraction process from data acquired from eight patients in pilot hospital. The features are acquired through a standardized protocol while clinicians provided also an extra feature, specifying the annotated level of a patient for the specific measurement. The protocol contained:

1) 6 minutes walking
2) 45 minutes supine position
3) 45 minutes standing position

IV. RESULTS

The Table II is the results table of the first dataset analysis. The Correlation-based Feature Subset Selection is being presented, because after our analysis, has the best accuracy between all applied feature selection algorithms. The first column in Table II displays the different classifiers that have been developed and applied. The second column contains the percentage of the correctly classified instances to the entire dataset, when all available and extracted attributes are being used in the input vector. Finally, the third column contains the percentage of the correctly classified instances, when a feature selection algorithm has been applied, the Correlation-based Feature Subset Selection. In this case only the attributes selected from the algorithm are being used in the vector that is inputted to the classifiers.

TABLE II

FIRST DATASET: CORRECTLY CLASSIFIED INSTANCES AND COMPARISON BETWEEN DIFFERENT APPLIED CLASSIFIERS

Applied Classifier (10 folds cross validation used)	Without Feature selection algorithm applied	Correlation-based Feature Subset Selection applied
Naïve Bayes	40 %	85 %
C4.5 Decision Tree	55 %	70 %
PARTial Decision Tree	35 %	60 %
Random Forest	40 %	70 %
Multilayer Perceptron	45 %	65 %
Support Vector Machine	50 %	65 %

The Table III is the results table of the second dataset analysis. The methodology and the applied algorithms are the same with those addressed for the first dataset.

TABLE III

SECOND DATASET: CORRECTLY CLASSIFIED INSTANCES AND COMPARISON BETWEEN DIFFERENT APPLIED CLASSIFIERS

Applied Classifier (10 folds cross validation used)	Without Feature selection algorithm applied	Correlation-based Feature Subset Selection applied
Naïve Bayes	90.00 %	82.50 %
C4.5 Decision Tree	97.50 %	97.50 %
PARTial Decision Tree	97.50 %	97.50 %
Random Forest	98.75 %	97.50 %
Multilayer Perceptron	98.75 %	93.75 %
Support Vector Machine	98.75 %	93.75 %

Table IV displays the results from the third dataset analysis. The methodology and the applied algorithms are the same with those addressed for the first dataset.

TABLE IV

THIRD DATASET: CORRECTLY CLASSIFIED INSTANCES AND COMPARISON BETWEEN DIFFERENT APPLIED CLASSIFIERS

Applied Classifier (10 folds cross validation used)	Without Feature selection algorithm applied	Correlation-based Feature Subset Selection applied
Naïve Bayes	73.07 %	73.08 %
C4.5 Decision Tree	61.53 %	65.38 %
PARTial Decision Tree	61.54 %	73.07 %
Random Forest	50.00 %	76.92 %
Multilayer Perceptron	73.08 %	96.15 %
Support Vector Machine	73.08 %	73.07 %

V. DISCUSSION

Analyzing the results we could clarify, in general, that the Naïve Bayes gives better result when is applied to the entire set of features. The applied Correlation-based Feature Subset Selection worsens the result. On the other hand, the Decision Tree algorithms give comparable results for both methodologies (applied feature selection algorithm and not applied). Finally, the Multilayer Perceptron and the Support Vector Machine give better results when the features are being filtered with the Correlation-based Feature Subset Selection algorithms in the first and third dataset, while the opposite stands for the second dataset.

When the second dataset is used and analyzed all methodologies give acceptable accuracy, while the Multilayer Perceptron and the Support Vector Machine exceed the 98% of correctly classified instances. This difference between the accuracy of the algorithms could be explained from the fact that the ratio features/instances varies between the three datasets. The second dataset, formed by 35 features and 80 instances is much more balanced and the algorithms are properly trained. In the third dataset, overtraining issues arise due to unbalanced ratio of features/instances and even when the feature selection algorithm is applied the problem isn't being eliminated.

VI. CONCLUSION

The developed system, a DSS module, is a part of an integrated platform for the monitoring of patients suffering from chronic diseases. The intelligent system has been implemented incorporating several classification methods and consists of a rule-based (expert) and a supervised classification system. In our preliminary analysis we have obtained positive results of the performance and the accuracy of the implemented system, which will be improved by the utilization of larger datasets.

REFERENCES

[1] G. Viegi, F. Pistelli , D.L. Sherrill, S. Maio, S. Baldacci and L. Carrozzi, "Definition, epidemiology and natural history of COPD," Eur Respir J; 30: pp. 993–1013, 2007.

[2] M. Tonelli, N. Wiebe, B. Culleton, A. House, C. Rabbat, M. Fok, F. McAlister and A.X. Garg, "Chronic Kidney Disease and Mortality Risk: A Systematic Review," J Am Soc Nephrol vol.17, pp. 2034-2047, 2006.

[3] C. Bellos, A. Papadopoulos, D. I. Fotiadis and R. Rosso, "An Intelligent System for Classification of Patients Suffering from Chronic Diseases" 32nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Buenos Aires, Argentina, August 31 - September 4, 2010.

[4] C. Bellos, A. Papadopoulos, R. Rosso και D. I. Fotiadis, "Extraction and Analysis of features acquired by wearable sensors network", 10th International Conference on Information Technology and Applications in Biomedicine, Corfu, Greece, 2-5 November 2010.

[5] M. Baig, Case-Based Reasoning, "An effective paradigm for providing diagnostic support for stroke patients", School of Computing, Queen's University, thesis report, 2008.

[6] N. Cristianini, J. Shawe-Taylor, "An introduction to support vector machines and other kernel-based learning methods", USA: Cambridge University Press; 2000.

[7] L. Brieman. "Random forests" Machine Learning Journal, 45, pp. 5-32,2001.

[8] Haykin, Simon, "Neural Networks: A Comprehensive Foundation", Prentice Hall, 1998.

[9] W. Liao and Q. Ji, "Pattern Recognition", 2009.

[10] F. V. Jensen and T. D. Nielsen, "Bayesian Networks and Decision Graphs", Springer Science and Business Media, 2007.

[11] N. Friedman, D. Geiger, M. Goldszmidt, "Bayesian Network Classifiers", Machine Learning, vol. 29, pp. 131-163, 1997.

[12] P. N. Tan, M. Steinbach, V. Kumar, "Introduction to data mining", Boston: Pearson Addison Wesley; 2006.

[13] E. Frank , I. H. Witten, "Generating Accurate Rule Sets Without Global Optimization", Proceedings of the Fifteenth International Conference on Machine Learning, p.144-151, July 24-27, 1998.

[14] L. Rokach, O. Maimon, "Data Mining with Decision Trees, Series in machine perception and artificial intelligence", Vol.69, pp.71-71, 2008.

[15] R. Quinlan, "C4.5", CA: Morgan Kauffman, 1993.

[16] E.E. Tripoliti et al., "A six stage approach for the diagnosis of Alzheimer's disease based on fMRI data", Journal of Biomedical Informatics , vol. 43, pp.307-320, 2010.

[17] D. Taniar, "Data Mining and Knowledge Discovery Technologies", part of the IGI Global series named Advances in Data Warehousing and Mining (ADWM) vol 2, 2007.