AUTOMATIC CREATION OF DECISION SUPPORT SYSTEMS: APPLICATION AND RESULTS IN THE CARDIOVASCULAR DISEASES DOMAIN

Markos G. Tsipouras, Themis P. Exarchos, Costas Papaloukas, Aris Bechlioulis, Anna Kotsia, Theodora Nanou, Christos Bazios, Yannis Antoniou, Dimitrios I. Fotiadis, Aikaterinh Naka, Lampros K. Michalis

Unit of Medical Technology and Intelligent Information Systems,
Department of Computer Science, University of Ioannina,
PO Box 1186, GR 451 10 Ioannina, GREECE
Tel:+302651097701, fax:+302651098889
Dept. of Cardiology, Medical School,
University of Ioannina, GR 45110, Ioannina, Greece

Email: markos@cs.uoi.gr,me01238@cc.uoi.gr, papalouk@cc.uoi.gr, md02798@yahoo.gr, hcsanna@cc.uoi.gr, nanou@cs.uoi.gr, chbazios@cc.uoi.gr, yanton@cc.uoi.gr, fotiadis@cs.uoi.gr, anaka@cc.uoi.gr, lmihalis@cc.uoi.gr

Abstract

In this paper we present the Decision Support Framework (DSF) of the NOESIS platform. NOESIS addresses wide scale integration and visual representation of medical intelligence in cardiology and aims at the development of a web-based personalized system with enhanced intelligence that supports health professionals in taking the best possible decision for diagnosis, prevention and treatment. The core of the NOESIS project is a set of Decision Support Systems (DSS), automatically generated from the DSF. Initially, the DSF has been employed to generate four DSSs, for the following cardiovascular sub-domains: (a) Ischaemic and (b) arrhythmic episode detection, (c) diagnosis of coronary artery disease (CAD) and (d) prediction of clinical restenosis in patients undergoing angioplasty.

Introduction

Cardiovascular diseases (CVD) constitute a primary pathology sector that leads to a significant number of deaths worldwide. Statistics related to CVD are disappointing. According to the World Health Organization (WHO), 16.6 million people around the globe die of CVD each year. The complexity and the uncertainty of cardiovascular diseases as well as their interconnection to other medical domains and health implications, make the diagnosis and the choice of treatment very difficult, especially for non-experienced cardiologists. The objective of the NOESIS system is to reduce uncertainty, in the complex domain of cardiovascular diseases, as much as possible. The NOESIS project provides an intelligent system that assists health professionals in promptly taking the best possible decision for diagnosis, prevention and treatment, by allowing a smooth transition from established medical knowledge to personal judgment. Under this scope, an intelligent environment that offers a wide range of services has been developed. Health professionals in hospitals, clinics and other health units might use NOESIS to enhance their knowledge on cardiovascular diseases, and support the process of medical decision. Medical schools and universities can use NOESIS as an educational tool for students and medical researchers. In addition, patients will also greatly benefit, since they receive timely and efficient healthcare services.

The core of the NOESIS project is a set of Decision Support Systems (DSSs), which support health professionals in taking the best possible decision in certain cardiovascular areas. These DSSs are automatically generated applying the DSF methodology to several different

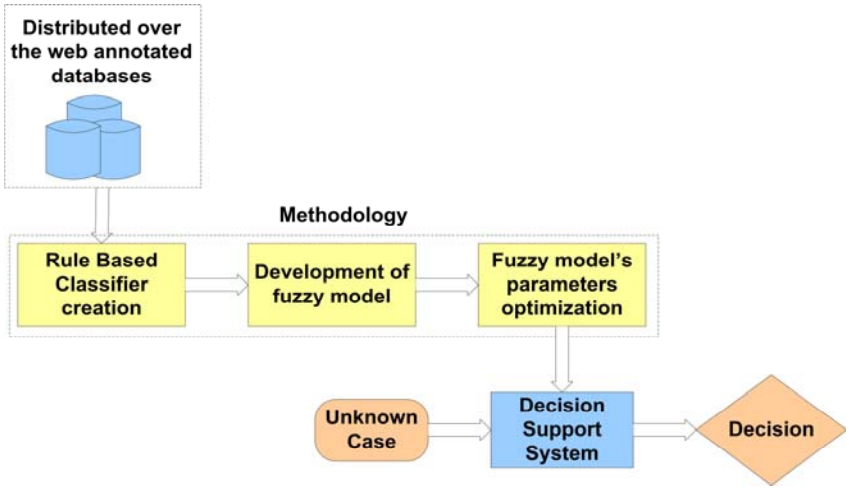
cardiovascular sub-domains. In order to use the DSF in a specific domain, an initial annotated dataset is required. This dataset is obtained from several resources, distributed over the web. The quality of the diagnosis, produced by the DSS, is analogous to the quality of this dataset. The DSF has been employed to generate four DSSs, for the following cardiovascular sub-domains: (a) Ischaemic and (b) arrhythmic episode detection, (c) diagnosis of coronary artery disease (CAD) and (d) prediction of clinical restenosis in patients undergoing angioplasty. For creating the DSS, the following elements are required: i) the symptoms which are described by the patient, ii) the clinical findings, described by the physician after his physical examination and iii) the test results (ECG). For each sub-domain the generated DSS and the corresponding results are presented.

Materials and Methods

The DSF is a three stage methodology [Exarchos et. al., 2005]: (a) creation of a rule-based classifier using medical knowledge or data mining techniques; (b) development of a fuzzy model; (c) optimization of the fuzzy model's parameters. In the first stage, a set of crisp rules is generated. This can be performed with two different approaches; the rules are provided by domain experts (knowledge-based approach) or are automatically generated from an annotated database using either association rule mining, or decision tree induction. In either case, a set of rules is created, in the form of a collection of "if…then…" crisp rules, represented in a disjunctive normal form (DNF), $\left(r_1 \vee r_2 \vee ...r_k\right)$, where r_i are the classification rules or disjuncts. Each rule is expressed as: $r_i : \left(Cond_i\right) \rightarrow y$, where y is the predicted class. The left-hand side of the rule is the rule antecedent or precondition. It contains a conjunction of feature tests: $Cond_i = \left(a_1\ op\ \theta_1\right) \wedge \left(a_2\ op\ \theta_2\right) \wedge ... \wedge \left(a_m\ op\ \theta_m\right)$, where $\left(a_j, \theta_j\right)$ is an feature-threshold pair and op is a comparison operator.

In the second stage, the crisp set of rules is transformed to a fuzzy set of rules using a fuzzy membership function instead of the crisp one, and S and T norms instead of the binary *AND* and *OR* operators. The sigmoid function is defined as: $g_s\left(a,\theta_1,\theta_2\right) = \left(1 + e^{\theta_1\left(\theta_2 - a\right)}\right)^{-1}$, while the minimum and maximum operators are used as *S* and *T* norms, respectively. According to these, the conjuncts are transformed to fuzzy ones as: $c_j^f\left(a_j,\theta_{1,j},\theta_{2,j}\right) = g_s\left(a_j,\theta_{1,j},\theta_{2,j}\right)$, and the crisp conditions as: $Cond_i^f\left(A,\Theta^f\right) = min\left\{c_{n_1}^f\left(a_{n_j},\theta_{1,n_j},\theta_{2,n_j}\right),...,c_{n_k}^f\left(a_{n_k},\theta_{1,n_k},\theta_{2,n_k}\right)\right\}$, where $\Theta^f = \left\{\theta_{1,1},\theta_{2,1},...,\theta_{1,n_f},\theta_{2,n_f}\right\}$ is a vector containing all parameters used in the fuzzy model. The general crisp rules are transformed to fuzzy ones as: $R_y^f\left(A,\Theta^f\right) = max\left\{Cond_{j_1}^f\left(A,\Theta^f\right),...,Cond_{j_n}^f\left(A,\Theta^f\right)\right\}$. These fuzzy general rules comprise the fuzzy model: $M^f\left(A,\Theta^f\right) = arg\ \max_{y=1,...,ny}\left(R_y^f\left(A,\Theta^f\right)\right)$, where n_y is the number of classes. For each feature vector A, the fuzzy general rule with the higher value defines its class.

Figure 1: The three-stage methodology used to create a Decision Support System



In the third stage, the fuzzy model $M^f\left(A,\Theta^f\right)$ is optimized with respect to its parameters Θ^f, using a training dataset (D_{train}). Having the confusion matrix defined as: $X_{M^f\left(A,\Theta^f\right),y} = \#\ of\ patterns\ in\ y\ classified\ to\ M^f\left(A,\Theta^f\right)$, the cost function, used for this purpose, is defined as: $F\left(\Theta,D_{train}\right)=\left|D_{train}\right|^{-1}\sum_{i=1}^{n_y}X_{i,i}$. The optimization method used is the HTMLSL, which is a stochastic algorithm based on MLSL. The above methodology is shown in Fig. 1.

Applications-Results

2.1 – Ischaemic Episode Detection

Myocardial ischemia is the condition of oxygen deprivation to the heart muscle and is accompanied by inadequate removal of metabolites due to reduced blood flow or perfusion. The detection of ischemic episodes can be very supportive to the physicians in the diagnosis of myocardial ischemia [Goldman, 1982, Papaloukas et. al., 2001]. The above presented framework (using association rule mining for rule extraction) has been evaluated for generating a DSS for ischemic beat detection. The dataset consists of 11 hours of two channel ECG recordings from the ESC ST-T Database [ESC ST-T DB, 1991]. The above recordings were preprocessed in order to remove noise like baseline wandering, A/C interference and EMG contamination. Then, five features were extracted from each cardiac beat: ST Segment Deviation, ST Segment Slope, ST Segment Area, T Wave amplitude and T Wave normal amplitude. In addition, a sixth feature, the patient's age was employed.

The recordings resulted in 76,989 cardiac beats, from which 1,936 were used for the generation of the ischemia DSS and 75,053 for testing it. The results, from the application of the framework employing only the first stage and then using all three stages, are presented in Table 1. After the detection of ischaemic beats, we followed the approach presented in [Papaloukas et. al., 2001], in order to detect the ischaemic episodes.

Table 1: Sensitivity and specificity results (%) for the ischaemic DSS

	DSF (1st stage)	DSF
Sensitivity	86.37	87.80
Specificity	90.43	92.46

2.2 – Arrhythmic Episode Detection

Arrhythmia can be defined as either an irregular single heartbeat (arrhythmic beat), or as an irregular group of heartbeats (arrhythmic episode). Arrhythmias can take place in a healthy heart and be of minimal consequence, but they may also indicate a serious problem that may lead to dangerous situations such as stroke or sudden cardiac death [Sandoe & Sigurd, 1991]. Therefore, automatic arrhythmia detection and classification is critical in clinical cardiology. For the classification of cardiac arrhythmias, the initial set of crisp rules was given by medical experts and then the final two stages of the framework were applied. The only feature used was the tachogram, extracted from ECG recordings with QRS detection. A three *RR* interval sliding window $\left[RR_1, RR_2, RR_3\right]$ was used to classify the middle *RR* interval $\left(RR_2\right)$ into one of the four categories: (1) ventricular flutter/fibrillation (VF), (2) premature ventricular contraction (PVC), (3) normal sinus rhythm (N) and (4) 2o heart block (BII). The dataset is $D = \left\{d^l, c^l\right\}$ with: $d^l = \left[RR_1, RR_2, RR_3\right]^l$, the l^{th} three RR interval window and c^l the class of the middle RR interval (RR_2). All beats from all records from the MIT-BIH [MIT-BIH, 1997] arrhythmia database were used to create and evaluate the arrhythmia DSS. The training dataset (D_{train}) is a randomly selected subset of D, containing equal number of patterns from each class (250) while the test dataset (D_{test}) consists of the remaining patterns of D. 20 different pairs of D_{train} and D_{test} are created. The mean values for sensitivity and specificity are presented in Table 2. After the classification of the beats, we followed the approach presented in [Tsipouras et. al., 2005], in order to classify the arrhythmic episodes.

Table 2: Sensitivity and specificity results (%) for the arrhythmic DSS.

	AF	PVC	N	BII
Sensitivity	99.05	81.79	95.62	98.98
Specificity	99.5	98.49	93.68	99.94

2.3 – Early Diagnosis of Coronary Artery Disease

Coronary artery disease (CAD) is the development of atherosclerotic plaques in the coronary arteries, resulting in coronary luminal narrowing and subsequently occlusion. CAD is the leading cause of death in western countries. Coronary angiography (CA), the “gold standard” method for the diagnosis of CAD, is an invasive and costly procedure [Vogel, 1997]. Therefore, a method able to predict non-invasively the presence of CAD would be of great clinical value. For the CAD diagnosis, the initial set of rules was generated from a decision tree. The dataset included 199 subjects suspected for CAD undergoing their first CA, and it was taken from the invasive cardiology department in the University Hospital of Ioannina. Patients with known CAD were excluded from the study. 89 of the subjects were normal subjects, and for the rest 110 the presence of CAD was confirmed by two experts. In order to characterize each subject, the 19 features shown in Table 3 were used.

Family history is defined as the presence of CAD in the father or brother of age<55 years or mother or sister of age<65 years. Hypertension was defined as systolic blood pressure (SBP) more than 140mmHg and/or diastolic blood pressure (DBP) more than 90mmHg or use of antihypertensive agents. Diabetes mellitus was defined as a fasting blood glucose concentration more than 126mg/dl or antihyperglycemic drug treatment. Current smoking was defined as having smoked the last cigarette less than a week before CA. Hyperlipidemia was defined as total cholesterol over 220mg/dl or use of lipid-lowering agents (statins or fibrates). Body mass index (BMI) was calculated as weight (kg) divided by the square of height (m2). Carotid – Femoral Pulse Wave Velocity (PWVcf) and Augmentation index (AIx)

were measured non-invasively using applanation tonometry, as indices of vascular stiffness. In order to assess the subjects status, concerning the existence or not of CAD, CA was performed by the Judkins technique. All coronary angiograms were visually assessed by two experienced angiographers and a consensus was reached. Significant CAD was defined as at least one 50% or greater diameter stenosis in at least one coronary artery vessel. The absence of CAD was defined as completely smooth epicardial coronary arteries.

The ten fold stratified cross validation method was used for evaluation. The procedure was applied to each fold, generating ten different crisp set of rules and fuzzy models. Both, the crisp set of rules and the final fuzzy model, have been evaluated in our dataset. Table 5 presents the average sensitivity and specificity. The overall accuracy of the crisp set of rules is 58%, while the average accuracy for the fuzzy DSS is 73%.

2.4 Prediction of clinical restenosis in patients undergoing angioplasty

Angioplasty with stent placement is currently taking the lead in the treatment of obstructive coronary artery disease (CAD). Restenosis, which occurs approximately in 12-60% of the patients within 6 months after intervention, depending on the patients' and procedural characteristics [Welt & Rogers, 2002], has been the main drawback to percutaneous-transluminal coronary angioplasty since its introduction. Therefore, identifying patients at increased risk for restenosis is important. For the prediction of clinical restenosis in patients undergoing angioplasty, the set of rules was generated from a decision tree. The dataset, was taken from the invasive cardiology department in the University Hospital of Ioannina and included 679 subjects.

Table 3: Features used in the CAD DSS

#	Feature	Units
1	Age	years
2	Sex	male(1), female(0)
3	Family History	yes(1), no(0)
4	Smoking	smoker (2), ex-smoker (1), non-smoker (0)
5	Diabetes	FBGC ≥ 126mg/dl (1) else (0)
6	Hypertension	DBP>90mmHg and/or SBP>140mmHg (1) else (0)
7	Hyperlipidemia	total cholesterol over 220mg/dl (1) else (0)
8	Creatinine	mg/dL
9	Glucose	mg/dL
10	Total Cholesterol	mg/dL
11	HDL	mg/dL
12	TRG	mg/dL
13	BMI	kg/ m2
14	Waist	cm
15	HR	bpm
16	SBP	mmHg
17	DBP	mmHg
18	PWVcf	m/sec
19	AIx	%

Table 4: Features used in the clinical restenosis DSS

#	Feature	Units
1	Age	years
2	Sex	male(1), female(0)
3	Family History	yes(1), no(0)
4	Smoking	smoker (1), non-smoker (0)
5	Diabetes	FBGC ≥ 126mg/dl (1) Else (0)
6	Hypertension	DBP>90mmHg and/or SBP>140mmHg (1) else (0)
7	Hyperlipidemia	total cholesterol over 220mg/dl (1) else (0)
8	CAD History	yes(1), no(0)
9	Prior PTCA	yes(1), no(0)
10	Prior CABG	yes(1), no(0)
11	Single Vessel Disease	yes(1), no(0)
12	Clinical Presentation	Unstable angina (1), Acute myocardial infarction (2), Stable angina (3)
13	Vessel Treated	Left anterior descending (1), Left circumflex (2), Right coronary artery (3), Left main (4), Bypass graft (5)
14	IIB/IIIA	yes(1), no(0)
15	Stent Type	Balloon (0), Stent (1)

In order to characterize the subjects, the 15 features shown in Table 4 were used. Family history, hypertension, diabetes mellitus, current smoking, hyperlipidemia are defined as above. Clinical presentation of CAD is classified as unstable angina, stable angina and acute myocardial infarction. The vessels treated with angioplasty are right coronary artery, left main coronary artery, left anterior descending artery, left circumflex artery and bypass grafts. The patients underwent either angioplasty with balloon or stenting and mainly bare metal stents. All patients were attended for a follow up time of at least 12 months. The composite end point of our study that is called clinical restenosis consists of cardiac death or a new non fatal myocardial infarction or a new revascularization attempt at the target vessel at 6 months of follow up after the angioplasty procedure.

All features that were used, except age, are binary or discrete valued and for this reason, the last two stages of the framework did not show any improvement. In addition, due to the imbalanced class distribution of our dataset (clinical restenosis was observed in 158 out of 679 subjects), the notion of cost sensitive learning (CSL) was introduced during decision tree induction. The two thirds of the dataset were used for training and one third for testing. Table 6 presents the sensitivity and specificity with and without the use of CSL.

Discussion

In the current study we introduced a novel methodology for automated generation of a DSS and its application to four cardiovascular sub-domains. The produced DSSs are evaluated and the obtained results demonstrate the usefulness of the generated DSSs as well as the overall methodology. The arrhythmia and ischaemia DSSs, which reported very high accuracy, comprise the ECG component of the NOESIS project that is a web application and performs ECG diagnosis in a web-based nature. The CAD DSS is a highly novel approach,

Table 5: Sensitivity and specificity results (%) for the CAD DSS

	DSF (1st stage)	DSF
Sensitivity	61.82	80.00
Specificity	53.93	65.17

concerning both the technical and the medical aspects. Coronary angiography (CA), the "golden standard" method for the diagnosis of CAD, is an invasive and costly procedure and thus cannot be used for screening of large populations. The proposed DSS requires only easily obtained data, i.e. the patient's history, routine blood tests and non-invasive assessment of vascular stiffness. The rule-based nature of the DSS makes the decision making process transparent. Furthermore, the use of CA for the initial annotation of the database is a great advantage, concerning the quality of our dataset. The fourth DSS provides an integrated system for decision support in clinical restenosis after angioplasty; there are no methods in the literature to addressing this problem. Finally, all the generated DSSs are web based applications and can be used remotely.

The generated systems could be used as tools to support the decisions of physicians in their daily clinical practice in hospitals or medical centres. Targeted users that will benefit by the use of the NOESIS platform are medical professionals that interact directly with the system and use its main functionality to support their personal, business needs, such as: cardiologists and cardiothoracic surgeons, general practitioners, medical students or researchers and pharmacologists. Secondary users are health professionals from other medical specialties e.g. radiologists, hematologists, paramedical staff such as nurses or

ambulatory personnel and marketing research departments in the health related industry. Further exploitation might focus on the application of the framework to other medical domains. This can be easily implemented, due to the fully automated nature of the framework, with the definition of other findings and appropriate diagnoses, provided by the experts. The evaluation of our framework in real clinical conditions is also of great interest.

Table 6: Sensitivity and specificity results (%) for the clinical restenosis DSS

	DSF (1st stage)	DSF (1st stage) with CSL
Sensitivity	40.00	64.43
Specificity	83.33	60.21

Acknowledgments

This research is part funded by the program "Heraklitos" of the Operational Program for Education and Initial Vocational Training of the Hellenic Ministry of Education and by the European Commission as part of the project NOESIS (IST-2002-507960).

References

Goldman, M.J. (1982). Principles of clinical electrocardiography, 11th ed. CA: LANGE Med. Pubs.

Papaloukas, C. Fotiadis, D.I. Liavas, A.P. Likas, A. and Michalis, L.K. (2001). A knowledge-based technique for automated detection of ischemic episodes in long duration electrocardiograms, Med. Biol. Eng. Comput., vol. 39, 105-112.

Sandoe, E. Sigurd, B. (1991). Arrhythmia-a guide to clinical electrocardiology. Bingen: Publishing Partners Verlags GmbH.

Vogel, R.A. (1997). Coronary Risk factors, Entothelial function and arteriosclerosis: A review, Clinical Cardiology, vol. 20, pp. 426-432.

Welt, F.G.P. Rogers, C. (2002). Inflammation and Restenosis in the Stent Era, Arterioscler. Thromb. Vasc. Biol., 22, 1769-1776.

Exarchos, T.P.et. al.(2005). A platform for wide scale integration and visual representation of medical intelligence in cardiology-The Decision Support Framework, in proc. 32nd CinC, France, 167-170.

European Society of Cardiology (1991). European ST-T database directory. Pisa: S.T.A.R.

Tsipouras, M.G. Fotiadis, D.I. and Sideris D. (2005). An arrhythmia classification system based on the RR interval signal., Artif. Intel. Med., 33, 237-250.

MIT-BIH arrhythmia database (1997). 3rd ed. Harvard-MIT Division of Health Sciences and Tech.