Personalization Algorithms Applied to Cardiovascular Disease Risk Assessment

*S. Paredes†, T. Marques§, T. Rocha†, P. de Carvalho‡, J. Henriques‡, J. Morais**

Abstract – Cardiovascular disease (CVD) is the major cause of death in the world. Clinical guidelines recommend the use of risk assessment tools (scores) to identify the CVD risk of each patient as the correct stratification of patients may significantly contribute to the optimization of the health care strategies.

This work further explores the personalization of CVD risk assessment, supported on the evidence that a specific CVD risk assessment tool may have good performance within a given group of patients and might perform poorly within other groups. Two main personalization methods based on the proper creation of groups of patients are presented: *i)* clustering patients approach; *ii)* similarity measures approach.

These two methodologies were validated in a Portuguese population (460 Acute Coronary Syndrome with non-ST segment elevation (ACS-NSTEMI) patients). The similarity measures approach had the best performance, achieving maximum values of sensitivity, specificity and geometric mean of, respectively, 77.7%, 63.2%, 69.7%. These values represent an enhancement in relation to the best performance obtained with current CVD risk assessment tools applied in clinical practice (78.5%, 53.2%, 64.4%).

I. Introduction

Cardiovascular disease (CVD) is caused by disorders of the heart and blood vessels and may include several specific conditions (coronary artery disease (CAD), heart failure, hypertension, stroke, etc.). CVD is the major cause of death in the world, representing only in Europe more than 47% of all deaths [1].

Prevention is the key to minimize this problem and should be implemented according to two different perspectives: *i)* lifestyle; *ii)* treatment. In fact, 77% of the disease burden in Europe is accounted for disorders related to lifestyle (unhealthy diet/obesity, physical inactivity, high blood pressure, smoking, etc.), while 80% of CAD could be prevented by maintaining healthy lifestyles [2]. Prevention should also be applied to the health care system, moving from reactive care towards preventive care and simultaneously transferring the care from the hospital to the patient's home. Here, health telemonitoring systems are very important as they allow the remote monitoring of patients who are in different locations away from the health care provider [3].

In this context, CVD risk assessment, i.e. the evaluation of the probability of occurrence of an event (death, myocardial infarction) given the patient's past and current exposure to risk factors, is important to the monitoring of each patient [4]. There are several risk assessment tools that were statistically validated and are applied in clinical practice. These tools calculate the probability of occurrence of a cardiovascular event within a certain period of time (months/years), considering different risk factors (e.g. age, sex, etc.). They can also differ in the endpoint/event (death, myocardial infarction, unstable angina, hospitalization), prevention type (primary/secondary) and patients' specific condition (e.g. diabetics) [5][6][7][8].

Clinical guidelines recommend the CVD risk assessment in order to aid the clinical decision as well as to contribute in making the patient more responsible for its own health. However, despite their clinical relevance, these tools present some important limitations [9].

Our recent research has addressed this problem through the development of different methodologies [9]. Firstly, a global framework (Bayesian model) was created directly from the fusion of the individual models parameters exploring the particular features of Bayesian inference mechanism. That methodology was based on two main hypotheses: *i)* it is possible to create a common representation of individual CVD risk assessment tools (naïve Bayes classifier); and *ii)* it is possible to combine individual models (the representations, naïve Bayes classifiers, of individual risk assessment tools). This allowed minimizing some of the identified problems such as dealing with missing risk factors, incorporation of additional clinical knowledge and clinical interpretability of the model. A personalization strategy, based on groups of patients, was also derived to address the problem of lack of performance. That methodology was supported on the evidence that a specific risk assessment tool may have a good performance within a given group of patients and might perform poorly within other groups [9].

The current work further explores this personalization concept, comparing the results obtained through two different strategies: *i)* clustering patients approach; *ii)* similarity measures approach.

The former involves the creation of groups of patients through a clustering algorithm that may be applied to the original data space or alternatively to a reduced dimension data space obtained through a dimension reduction technique. Then, the most appropriate CVD risk assessment tool (tool that presents the best performance) is identified for each group of patients. The latter also considers groups of patients but in a different

This work was supported by CardioRisk (PTDC /EEI-SII/2002/2012), iCIS (CENTRO-07- ST24 – FEDER – 002003) and CISUC. (Center for Informatics and Systems of University of Coimbra).

† Instituto Politécnico de Coimbra, Departamento de Engenharia Informática e de Sistemas, Portugal, {sparedes@isec.pt, teresa@isec.pt}.

§ Departamento de Engenharia Informática, Universidade de Coimbra, Coimbra, Portugal, {tmarques@student.dei.uc.pt}

‡ CISUC, Departamento de Engenharia Informática, Universidade de Coimbra, Coimbra, Portugal, {carvalho@dei.uc.pt, jh@dei.uc.pt}.

*Serviço de Cardiologia, Leiria Pombal Hospital Centre, Portugal, {joaomorais@hsaleiria.min-saude.pt}.

perspective. The patients that are correctly classified by a specific CVD risk tool form a group. A new patient is assigned to the group that its nearest neighbor belongs to. This assignment is based on similarity measures.

The validation phase was supported by a real patient testing dataset obtained in the Santa Cruz Hospital, Lisbon/Portugal, that comprises N=460 ACS-NSTEMI patients.

The paper is organized as follows: in section II an outline of the methodologies is presented. In section III the results obtained with the two personalization strategies are presented. Section IV summarizes the main conclusions.

II. METHODOLOGY

The proposed methods rely on the creation of groups of patients in order to achieve the personalization of CVD risk assessment. However, the two approaches create these groups according to two divergent perspectives.

Clustering algorithms are unsupervised learning algorithms, therefore the identification of groups of patients is exclusively based on the values of the respective risk factors. Similarity measures approach implements a different concept of group of patients, as it assumes that the patients correctly classified by a specific risk assessment tool belong to the same group. The new instances are assigned to the different groups using similarity measures.

Both situations require the use of distance metrics that should be selected according to the type/nature of data [10]. The selection of the best distance metric to identify similarities among patients is not a trivial step, thus several distance metrics (Figure 1) are tested to identify the one that originates the best results.

Euclidean
$$d(\mathbf{x},\mathbf{y}) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

Mixed
$$d(\mathbf{x},\mathbf{y}) = \frac{\sum_{i=1}^{n} \begin{cases} x_i = y_i & x_i, y_i \ nominal \\ \frac{|x_i - y_i|}{\max_i - \min_i} & x_i, y_i \ interval \end{cases}}{n}$$

Hamming
$$d(\mathbf{x},\mathbf{y}) = \#(x_i \neq y_i)$$

Jaccard
$$d(\mathbf{x},\mathbf{y}) = \frac{\#[(x_i = y_i) \cap ((x_i \neq 0) \cup (y_i \neq 0))]}{\#[(x_i \neq 0) \cup (y_i \neq 0)]}$$

$\mathbf{x}, \mathbf{y}$ data vectors; xi attribute i of instance $\mathbf{x}$; yi attribute i of instance $\mathbf{y}$; n number of attributes; $maxi$ maximum value of attribute i; $mini$ minimum value of attribute i

Figure 1 - Distance metrics.

Due to the different nature of attributes/risk factors (interval-scaled; ordinal; binary variables) [11] and according to the distance metric to be used, a discretization or a normalization step must be performed. Normalization is performed according to z-score (1):

$$z = \frac{x - \mu}{\sigma} \quad (1)$$

where μ is the mean of population and σ is the respective standard deviation. Discretization is implemented similarly to the equal width discretization method (EWD), being the value of the attribute rounded to the nearest power of ten.

A. Clustering Patients Approach

Figure 2 presents the two main phases of clustering patients approach: *i)* training process ; *ii)* classification.

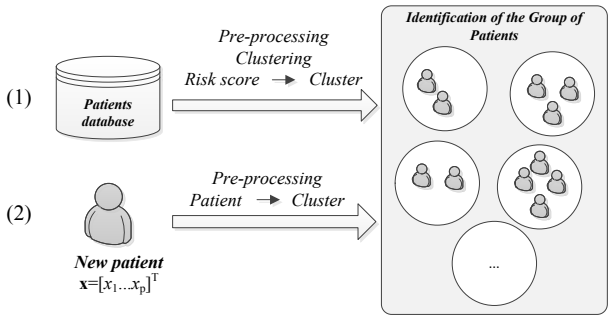

Figure 2- Clustering Patients Approach

The training process involves the creation of a set of clusters that allows the identification of the best CVD risk assessment tool to classify a new patient.

Initially data is pre-processed (normalized/discretized) according to the distance metric used to the clusters creation.

The second step consists of a clustering procedure, where groups of patients are created based on the respective values of the considered risk factors. In this step two well-known dimensionality reduction techniques (linear technique: PCA, non-linear technique: Kernel PCA) were applied [12]. The dimensionality reduction methods attempted to improve the results as a high number of variables (dimensions) may decrease the efficiency (performance/accuracy) of data mining algorithms [12]. Features selection, an alternative strategy to overcome the high dimensionality problem, was also tested. Actually, high dimensional data may contain features that are irrelevant/redundant for the classifier performance. A correct assignment of weights can eliminate or reduce the importance of those features [13]. Feature selection comprises two main approaches: *i)* filter model where a subset of features is chosen without any data obtained from the classifier; *ii)* wrapper model where the quality of selection is guided by the classifier [13]. Three filter models (Gini index, Relief-F algorithm and Fast Correlation Based Filter) and one wrapper model (random search) were tested.

Patients are grouped based on the respective risk factors. Thus, the goal is to apply a clustering algorithm to $X_{n \times N}$ where n is the number of risk factors, and N is the number of patients in the training dataset in order to create K disjoint groups (clusters) $G = \{G_1, ..., G_K\}$ of patients. Here, it is important to test different dimensions (n) of the data space.

The clusters are created through the subtractive clustering, which is a density based algorithm, since it groups data instances according to their density (a data point $\mathbf{x}_i$ will have a high density value if it has many neighbouring data points). The first cluster center is the one that presents the highest

density value. The density values are updated and the process iterates to find the next cluster center until an adequate number of clusters is identified [14]. Subtractive clustering was selected as it assures ability to deal with large number of instances, capacity to group mixed attributes, insensitivity to the order of attributes and handling of outliers [14].

After the clusters creation, CVD risk assessment tools are assigned to the several clusters based on the respective performance (SE: sensitivity; SP: specificity; G_{mean}: geometric mean). Among the clusters with $G_{mean} > 0.5$, the risk assessment tool with the highest G_{mean} is selected. In those clusters where $SE > 0.5$ and the $G_{mean} < 0.5$ ($G_{mean} = 0$ if the cluster only contains instances of one class) the selection is based on the highest value of SE. Otherwise it is supported in the specificity's value.

Therefore, a new patient is assigned to a specific cluster being classified by the CVD risk assessment tool with the best performance in that cluster.

However, there are some difficulties related with this clustering methodology. The proper creation of representative groups of patients requires a large training dataset, which may be difficult/expensive to obtain. Additionally, the clustering process is complex, as it involves finding similarities between instances and creating groups with the appropriate dimension (i.e. if the cluster is too big it may not provide differentiation among the performance of the several risk assessment tools; if the cluster is too small it makes it impossible to apply the concept of patients grouping).

B. Similarity Measures Approach

This methodology proposes a simpler strategy to form groups of patients. Unlike, the clustering approach that is based on an unsupervised learning algorithm, here the groups are created according to the patients classification as shown in Figure 3.

The classification of a new patient is based on similarity measures. Therefore, if a new patient is closest to one that is correctly classified by a risk score, it is probable that the same risk score will also be able to classify it accurately. In this way, the groups of patients are formed by the patients correctly classified by each score. If a patient is not correctly classified by any of the individual CDV risk assessment tools, it is assigned to a group that is classified by the CVD risk tool with highest sensitivity when applied to the entire training dataset.

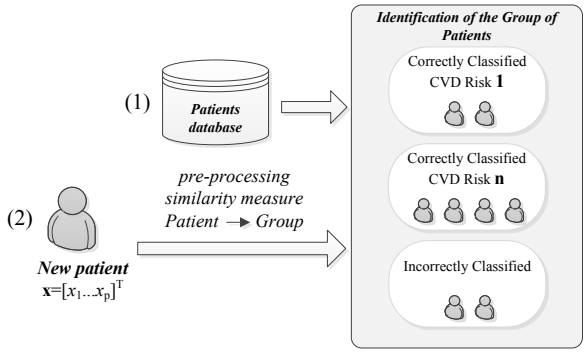

Figure 3- Similarity Measures Approach

Nevertheless, the identification of the closest patient is not obvious. In order to achieve the best performance, several distances must be considered (Figure 1).

Similarly to the clustering approach, dimension reduction techniques were applied as well as feature selection algorithms, namely a random search (wrapper algorithm). In this way the similarity of patients can be refined, as risk factors have different levels of relevance to calculate that similarity.

C. Validation

The training and testing data sets were directly obtained from a real patient dataset. The two methodologies were validated with 10 fold-cross validation and 30 runs. Some statistical tests (Friedman's ANOVA complemented with Bonferroni correction) were also applied. This strategy aimed to reinforce the validation conclusions.

III. Results

A. Training and Testing Datasets

Training and testing datasets were obtained from a real patient dataset from the Santa Cruz hospital, Lisbon, Portugal. This dataset contains data from N=460 consecutive patients that were admitted in the Santa Cruz Hospital, Lisbon, with ACS-NSTEMI between March 1999 and July 2001. The event rate of combined endpoint (death/myocardial infarction) is 7.2% (33 events).

B. CVD Risk Assessment Tools

The proposed methodologies were validated considering three well known CVD risk assessment tools (GRACE [6], TIMI [7], PURSUIT [8]), developed for short-term (1 month) in secondary prevention (CAD patients) (Table I).

TABLE I
SHORT-TERM RISK ASSESSMENT MODELS

Model	Risk Factors
GRACE [6]	Age, SBP, CAA HR, Cr, STD, ECM, CHF
TIMI [7]	Age, STD, ECM, KCAD, AS, AG, RF
PURSUIT [8]	Age, Sex, SBP, CCS, HR, STD, ERL, HF

Cr-Creatinine, **HR** – Heart Rate, **CAA** – Cardiac Arrest at Admission, **CHF** – Congestive Heart Failure, **STD** - ST Segment. Depression, **ECE** - Elevated Cardiac Enzymes, **KCAD**- Known CAD, **ERL** – Enrolment (MI/UA), **HF** –Heart Failure, **CCS** – Angina classification, **AS** - Use of aspirin in the previous 7 days, **AG** - 2 or more angina events in past 24 hrs, **RF** - 3 or more cardiac risk factors

Table II presents the performances obtained with the CVD risk tools in the considered dataset.

TABLE II
CVD RISK ASSESSMENT TOOLS PERFORMANCE – SANTA CRUZ, (DEATH/MI)

%	GRACE	PURSUIT	TIMI
SE	78.56 ± 0.3	63.91 ± 1.2	69.25 ± 1.3
SP	53.18 ± 0.01	56.01 ± 0.03	43.79 ± 0.04
Gmean	64.42 ± 0.3	58.64 ± 1.3	53.09 ± 1.0

GRACE achieved the best performance obtaining the highest sensitivity and G_{mean}.

C. Clustering Patients Approach

The clusters' creation had to be performed iteratively. In fact subtractive clustering (density based) produces different results according to the respective parameters.
Table III presents the best results achieved with different neighborhood radius and distance metrics.

TABLE III
CLUSTERING PATIENTS - DISTANCE METRICS/NEIGHBORHOOD RADIUS

%	Euclidean (0.1)	Mixed (0.5)	Hamming (0.5)	Jaccard (0.5)
SE	76.07 ± 2.3	78.14 ± 1.35	73.64 ± 3.01	76.19 ± 3.07
SP	57.83 ± 0.84	58.15 ± 0.83	53.22 ± 1.33	53.52 ± 0. 45
G_{mean}	65.93 ± 1.17	67.15 ± 0.77	62.06 ± 1.73	63.45 ± 1.49

Mixed distance achieved the best performance. It had a SE similar to GRACE while improved the SP and consequently the G_{mean}. Dimension reduction techniques (PCA, KPCA) were applied but deteriorated the results. Some feature selection methods were also tested but without good results.

D. Similarity Measures Approach

Table IV presents the results obtained with this strategy without assigning weights to the risk factors.

TABLE IV
SIMILARITY MEASURES APPROACH

%	Euclidean	Mixed	Hamming	Jaccard
SE	66.76 ± 1.7	66.19 ± 1.63	72.33 ± 1.8	69.54 ± 1.57
SP	64.64 ± 0.51	66.13 ± 0.37	60.54 ± 0.41	62.24 ± 0. 51
G_{mean}	64.68 ± 1.4	65.08 ± 1.30	65.52 ± 1.21	64.94 ± 1.33

The performance, namely the sensitivity value, has been reduced in all the test cases, which is unacceptable. As a result a weighted distance was implemented through random search (Table V, Table VI).

TABLE V
RANDOM SEARCH - WEIGHTS

n.	Weights
3	[1.0,0.2,0.6,0.5,0.8,0.8,0.4,0.6,0.5,0.2,0.9,0.1,0.4,0.8]
4	[1.0,0.9,0.3,0.6,0.8,0.9,0.9,0.5,0.9,0.7,0.3,0.2,0.2,0.8]
9	[0.5,0.1,0.8,0.5,1.0,0.4,0.5,0.4,0.9,0.3,0.2,1.0,0.3,0.8]

TABLE VI
SIMILARITY MEASURES APPROACH – WEIGHTED DISTANCE

%	Euclidean weights n. 3	Mixed weights n. 4	Hamming weights n. 9	Jaccard weights n. 9
SE	71.79 ± 1.78	71.48 ± 1.91	77.33 ± 1.38	77.67 ± 1.72
SP	65.86 ± 0.54	67.47 ± 0.52	61.86 ± 0.44	63.16 ± 0.38
G_{mean}	68.20 ± 0.98	68.86 ± 1.07	69.09 ± 1.21	69.74 ± 0.87

The results show that the Euclidean and Mixed distances were able to improve the specificity, but reduced the original sensitivity's value. The Hamming and Jaccard distances did not improve so much the specificity, however, the sensitivity was also not as decreased as in the other two distances. Actually, with these two distances and the proper set of weights, the similarity measures approach achieved a sensitivity's value similar to GRACE tool and simultaneously increased the value of specificity significantly. Statistical significance tests reinforced these results.

IV. FUTURE WORK AND CONCLUSIONS

Considering the obtained results it is possible to affirm that personalization of the CVD risk assessment, based on groups of patients, can be a valid contribution to improve health care. Particularly the similarity measures approach, achieved very interesting results, as it retained the GRACE sensitivity value while significantly increasing the specificity value, which resulted in a better global performance.
Some developments of the proposed methodologies can be implemented, e.g. replacement of the random search by a solution based on genetic algorithms.
Additionally and in spite of these promising results, further tests with larger and more balanced datasets would be very useful to strengthen these conclusions.

V. REFERENCES

[1] Nichols M. *et. al*., "Cardiovascular disease in Europe: epidemiological update", EHJ, Vol. 34, pp. 3028-3034, 2013.
[2] Boye N. *et. al*., "PREVE White Paper – ICT Research Directions in Disease Prevention", FP7 – 248197, 2010.
[3] Reiter, N. *et. al*., "HeartCycle: Compliance and Effectiveness in HF and CAD Closed-Loop Management", Proc. of the 31st Annual International Conference of the IEEE EMBS, pp. 299-302, 2009.
[4] Graham, I. *et. al*., "Guidelines on preventing cardiovascular disease in clinical practice: executive summary", European Heart Journal, Vol.28, pp. 2375 – 2414, 2007.
[5] Siontis G., "Comparisons of established risk prediction models for cardiovascular disease: systematic review", BMJ 2012; 344:e3318.
[6] Tang. E, *et. al*., "Global Registry of Acute Coronary Events (GRACE) hospital discharge risk scores accurately predicts long term mortality post-acute coronary syndrome", AHJ, Vol. 154, pp. 29-35, 2007.
[7] Antman, E. *et. al*., "The TIMI risk score for Unstable Angina/ Non-St Elevation MI - A method for Prognostication and Therapeutic Decision Making", JAMA ,Vol. 284, pp. 835-842, 2000.
[8] Boersma E, *et. al*., "Predictors of outcome in patients with acute coronary syndromes without persistent ST-segment elevation. Results from an international trial of 9461 patients", Circulation Vol. 101, pp. 2557–2567, 2000.
[9] S. Paredes *et. al*., "Cardiovascular Disease Risk Assessment Innovative approaches developed in HeartCycle project", 35th Annual International IEEE EMBS, pp. 6980-3, 2013.
[10] Choi, S. *et al.,* "A Survey of Binary Similarity and Distance Measures." Cybernetics and Informatics, Vol.8, pp. 43-48, 2010.
[11] P. Andritsos *et al.*, \Data clustering techniques," Toronto, University of Toronto, Dep. of Computer Science, Technical Report CSRG-443, 2002.
[12] Maaten L., Postma E., Herik H., "Dimensionality reduction: A comparative review," Journal of Machine Learning Research, vol. 10, pp. 1–41, 2009.
[13] Molina L., Belanche L., and Nebot A., "Feature selection algorithms: A survey and experimental evaluation,". Proceedings of ICDM International Conference, pp. 306–313, IEEE, 2002.
[14] Han J. "Data Mining: Concepts and Techniques, 3rd edition". ISBN: 978-0123814791, Morgan Kaufmann, 2011.