Public Health and Informatics
J. Mantas et al. (Eds.)

doi:10.3233/SHTI210111

A Constructive Fuzzy Representation Model for Heart Data Classification

Michael D. VASILAKAKIS[a], Dimitris K. IAKOVIDIS[a, 1] and George KOULAOUZIDIS[b]
[a] *Dept of Computer Science and Biomedical Informatics, University of Thessaly,* Lamia, Greece
[b] *Department of Cardiology, Stepping Hill Hospital Stockport, United Kingdom*

Abstract. The early detection of Heart Disease (HD) and the prediction of Heart Failure (HF) via telemonitoring and can contribute to the reduction of patients' mortality and morbidity as well as to the reduction of respective treatment costs. In this study we propose a novel classification model based on fuzzy logic applied in the context of HD detection and HF prediction. The proposed model considers that data can be represented by fuzzy phrases constructed from fuzzy words, which are fuzzy sets derived from data. Advantages of this approach include the robustness of data classification, as well as an intuitive way for feature selection. The accuracy of the proposed model is investigated on real home telemonitoring data and a publicly available dataset from UCI.

Keywords. Data representation, Feature extraction, Classification, Health, Cardiology, Telemonitoring

1. Introduction

Heart Disease (HD) and Heart Failure (HF) are complex clinical problems related to many different pathological factors making their diagnosis a complicated task. HF often occurs as a complication of HD. They are both associated with high mortality rate and frequent hospital admissions. Early treatment and regular follow up, usually by telemonitoring (TM), are key strategies to address them [1]. The diagnosis of these heart problems depends on diverse features; therefore, it is important to be able to distinguish and select appropriate features as markers to provide the highest possible diagnostic accuracy. HD diagnosis and HF prediction have been investigated in several studies using machine learning approaches [2, 3]. Among the studies investigating the detection of HD, a recent one [4], was based on a supervised feature selection and classification method, considering the bounded sum of weighted fuzzy membership functions. In [5] the optimization of the granularization of the feature space has been considered in a more general fuzzy classification framework. Fewer works have been performed with respect to HF prediction. A recent methodology was presented in [6] based on features obtained by Multi-Resolution Analysis (MRA) of TM signals.

In this paper, we propose a novel Constructive Fuzzy Representation Model (CFRM), aiming to enhance the expressivity of the collected heart features, and

[1] Corresponding Author, Dimitris K. Iakovidis, E-mail: diakovidis@uth.gr

consequently to improve heart data classification. Since Zadeh [7] established the

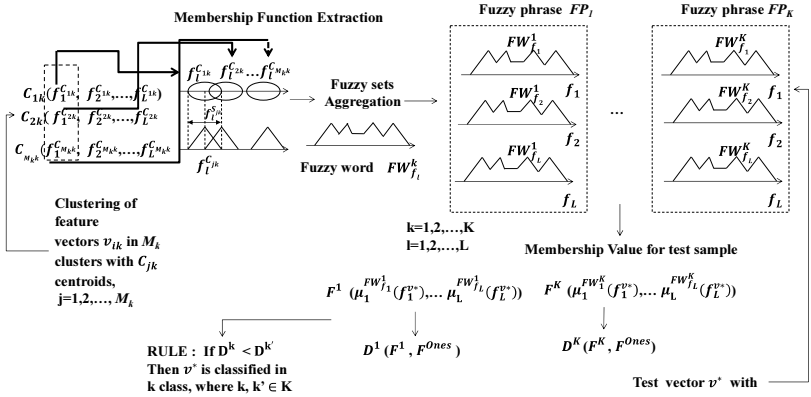

Figure 1. In the training phase the features of each vector are used for the extraction of the fuzzy phrases. In the test phase the features of the unknown test sample are described based on the membership calculated from its features to the fuzzy words. Finally, the test sample is classified based on the adopted decision rule.

foundations of the fuzzy sets theory, these have been applied in a variety of research domains for various classification and pattern recognition models, including medical diagnosis [8]. The proposed model follows an intuitive approach for the description of data that resembles the way humans use specific vocabularies of words for the description of real-world concepts. In addition, a feature selection methodology, embedded into CFRM is proposed.

2. Methods

2.1. CFRM Training and Test Phases

The proposed CFRM methodology is illustrated in Figure 1. Let K be the number of different classes of the classification problem under investigation, and N_K be the number of training feature vectors v_{ik} from each class in the training phase. Every feature vector is composed of different features. Let $v_{ik}(f_1^{v_{ik}}, f_2^{v_{ik}}, \dots, f_L^{v_{ik}})$ be an L-dimensional feature vector extracted from a training sample, with features $f_l^{v_{ik}}$, l=1,2,…, L, k=1,2,...,K, where K is the number of classes, and i=1,2...,N_k, where N_k is the number of training samples per class k. In the training phase, CFRM applies a clustering algorithm to cluster the feature vectors $v_{ik}, i = 1,2..N_k$, into a set of $M_K < N_K$ clusters. Every cluster has a centroid C_{jk}, which has the form $C_{jk}(f_1^{C_{jk}}, f_2^{C_{jk}}, \dots, f_L^{C_{jk}}), j = 1,2, \dots, M_K$. Each feature $f_l^{C_{jk}}$, l=1,2,…,L of the centroid C_{jk} of the j^{th} cluster represents a centroid of the features $f_l^{v_{ik}}$ in l^{th} dimension with a respective standard deviation $f_l^{S_{jk}}$ of the features $f_l^{v_{ik}}$.

After the computation of the centroid coordinates $f_l^{C_{jk}}$ and their standard deviations $f_l^{S_{jk}}$ of the features $f_l^{v_{ik}}$ in l^{th} dimension, a fuzzy set can be defined with a respective membership function having the form $\mu_l^{\left(f_l^{C_{jk}}\right)}\left(f_l^{v_{ik}}\right)$. The fuzzy sets of a class k, which are defined according to the procedure, are aggregated using the union operation,

$$FW_{f_l}^k = \bigcup_{j=1}^{M_k} \mu_l^{\left(f_l^{c_{jk}}\right)}\left(f_l^{v_{ik}}\right) \quad (1)$$

for l=1,2,…, L, k=1,2,..., K, and i=1,2...,N_k. The new fuzzy sets $FW_{f_l}^k$ defined by this aggregation operation are considered as fuzzy words. Let $\mu_l^{FW_{f_l}^k}$ be the aggregated membership function of each fuzzy word $FW_{f_l}^k$. A feature $f_l^{v_{ik}}$ of a feature vector v_{ik} is a member of fuzzy word $FW_{f_l}^k$ if

$$0 < \mu_l^{FW_{f_l}^k}\left(f_l^{v_{ik}}\right) \le 1 \quad (2)$$

for l=1,2,…,L, k=1,2,...,K, and i=1,2...,N_k. A set of fuzzy words $FP_k = \{FW_{f_1}^k, FW_{f_2}^k, \ldots, FW_{f_L}^k\}$ defines a *fuzzy phrase*, which is representative for class k.

During the test phase, let v^* be an unknown sample that is represented with the feature vector $v^*\left(f_1^{v^*}, f_2^{v^*}, \ldots, f_L^{v^*}\right)$. For each feature $f_l^{v^*}$, l=1,2,…,L, the respective membership to the fuzzy sets $FW_{f_l}^k$, k=1,2,...,K. is computed. The overall membership of the feature vector v^* to a class k is represented by a membership vector $F^k(\mu_1^{FW_{f_1}^k}, \mu_2^{FW_{f_2}^k}, \ldots \mu_{\mathrm{L}}^{FW_{f_L}^k})$ where each feature $\mu_l^{FW_{f_l}^k}$ represents a membership to the fuzzy word $FW_{f_l}^k$. Consequently, the test sample with feature vector $v^*\left(f_1^{v^*}, f_2^{v^*}, \ldots, f_L^{v^*}\right)$ is described by F^k, k=1,2,...,K membership vectors, one for each class.

In order to classify the test sample with feature vector v^* to a class, we adopt a rule based on the distance of each membership vector $F^k(\mu_1^{FW_{f_1}^k}, \mu_2^{FW_{f_2}^k}, \ldots \mu_{\mathrm{L}}^{FW_{f_L}^k})$ k=1,2,...,K from a membership vector with components equal to one. Since the maximum value of a membership is equal to one and that the minimum is equal to zero, the ideal case of a sample from the class k is to have all the features of the membership vector F^k equal to one. Thus, a $F^{ones}(f_1^{ones}, f_2^{ones}, \ldots, f_L^{ones})$ membership vector is defined to represent the ideal case, where each feature $f_l^{ones}, l = 1 \ldots L$, is $f_l^{ones} = 1$. Let $D^k(F^k, F^{ones})$ be the distances of F^k, k=1,2,..., K from F^{ones}, and $D^{k'}$ $(F^{k'}, F^{ones})$ be the distances of $F^{k'}$ and F^{ones}, where $k, k' \in K$ and $k \neq k'$. The rule "*If D^k<$D^{k'}$ then v^* classified in class k*" can be used for classification of v^*.

CFRM can embed feature selection to reduce the complexity of the classification task and to identify the most informative features within a dataset. The CFRM-based Feature Selection (CFRM-FS) is based on the information, derived from the fuzzy words FW_l^k, defined during the training phase. More specifically, Figure 2 shows that each class k is represented by a fuzzy phrase FP_k, which is a set of fuzzy words FW_l^k. These fuzzy words are fuzzy sets that have been produced by clustering of the feature vectors v_{ik}. The overlap between two fuzzy words indicates that these words carry redundant information; therefore, the respective feature $f_l^{v_{ik}}$ can be considered as less important. For example, the fuzzy word illustrated in Figure 2(b) is expected to be less informative than the fuzzy word illustrated in Figure 2(a). The overlap h of the fuzzy words FW_l^k and $FW_l^{k\prime}$ of the classes k and k' respectively, is estimated by the intersection divided by the union of these fuzzy words. However, h can be considered only as a weak indicator of redundancy, since it results from a non-deterministic clustering procedure (*e.g.*, clustering algorithms, such as the k-means, usually depend on a random initialization and arbitrarily determined parameters, such as the number of clusters). A stronger redundancy indicator can be obtained by aggregation of multiple overlap observations from multiple executions of the clustering algorithm. Based on this approach, a feature is selected as informative if and only if all overlap observations of the respective words, are low.

(a) (b)

Figure 2. (a) Fuzzy words $FW_{f_l}^{k}$ and $FW_{f_l}^{k\prime}$ of two different classes k and k' with lower overlap. (b) Fuzzy words $FW_{f_l}^{k}$ and $FW_{f_l}^{k\prime}$ of two different classes k and k' that have high overlap.

3. Results

CFRM is evaluated on two case studies using 10-fold cross-validation. The first one addresses the diagnosis of HD, and the second one addresses the telemonitoring of HF patients for hospitalization prediction, using daily collected physiological data. The HD Dataset (HDD) is from the Cleveland Clinic Foundation, it contains 270 samples and it is publicly available [9]. The classification problem under study with respect to this dataset is the absence (150 samples) or presence (120 samples) of HD, based on 13 features (Age, Sex, Chest pain type, Resting blood pressure, Serum cholesterol in mg/dl, Fasting blood sugar > 120 mg/dl, Resting electrocardiographic (ECG) results, Maximum heart rate achieved, Exercise, Oldpeak = ST depression in the ECG signal induced by exercise relative to rest, The slope of the peak exercise ST segment, Number of major vessels, Defect type). HF Dataset (HFD) is based on retrospective telemonitoring data from 308 patients in Kingston-upon-Hull [6]. The dataset is fully anonymized and, in the study, the cases for which death has been reported are 6.5% and considers only the information extracted from the monitored physiological signals heart rate (HR), systolic blood pressure (SBP), diastolic blood pressure (DBP), and body weight (BW). The MRA features extracted from HFD correspond to time-intervals of 4-day patient monitoring. The dataset is highly imbalanced, with only 0.2% of the vectors representing Worsening HF (WHF) precursor patterns, i.e., vectors corresponding to HFD. This makes the classification task even more challenging. Another challenge in this dataset is that in one fourth of the patients' data there are missing values during consecutive 4 days prior to HFH, respectively. This is due to low compliance of the patients to the use of TM.

Table 1. HDD Comparative Results

Method	Evaluation metrics			
	AUC	**Accuracy**	**Sensitivity**	**Specificity**
CFRM	0.81	0.82	**0.80**	0.87
CFRM-FS	**0.88**	**0.85**	0.71	**0.94**
Lee [4]	N/A	0.82	N/A	N/A
Hu *et al*. [5]	0.74	0.75	N/A	N/A

N/A: Not Available

Table 2. HFD Comparative Results

Method	Evaluation metrics			
	AUC	**Accuracy**	**Sensitivity**	**Specificity**
CFRM	**0.80**	**0.93**	**0.50**	**0.94**
CFRM-FS	0.78	0.88	0.38	0.88
MRA [6]	0.76	N/A	0.47	0.96
MRA BW [6]	0.75	N/A	0.38	0.98
MRA BW, DBP [6]	0.77	N//A	0.48	0.96

N/A: Not Available

In the case of HF prediction, CFRM and CFRM-FS outperform the state-of-the-art method proposed in [6] (Table 2), while it is worth mentioning that the features selected using CFRM-FS are the same with those selected in [6]. It can also be noticed that CFRM

performs better than CFRM-FS and this could be attributed to the fact that HFD is particularly imbalanced.

4. Discussion and Conclusion

In this study we proposed a novel model to enhance the expressivity of feature for better data classification, named CFRM and it was extended using a novel feature selection methodology, named CFRM-FS. The performance of both CFRM and CFRM-FS were investigated and evaluated. Both approaches resulted in a better or comparable performance from the previously reported methodologies for detection of HD and prediction of HF. Overall the main benefit of the proposed model is its capability to effectively perform both a selection of the most informative features in a dataset, and data classification, in a computationally efficient way. The CFRM-based classification method is still in an early stage; however, its generality and the results obtained are promising also for other applications in the domain of medical informatics, whereas due to its intuitive interpretation it can be considered as a basis for developing interpretable machine learning-based medical decision support models. Further investigation is required to fully explore all its potentials. Directions for further research include alternatives rules for decision making, and systematic evaluation of its robustness to noise and the presence of missing values.

Acknowledgments

We acknowledge support of this work by the project “Smart Tourist” (MIS 5047243) which is implemented under the Action “Reinforcement of the Research and Innovation Infrastructure”, funded by the Operational Programme "Competitiveness, Entrepreneurship and Innovation" (NSRF 2014-2020) and co-financed by Greece and the European Union (European Regional Development Fund).

References

[1] Timmis A, Townsend N, Gale C, et al. European Society of Cardiology: cardiovascular disease statistics 2017. European heart journal 2018; 39: 508–579.
[2] Shah M, Zimmer R, Kollefrath M, et al. Digital Technologies in Heart Failure Management. Current Cardiovascular Risk Reports 2020; 14: 1–8.
[3] Khan Y, Qamar U, Yousaf N, et al. Machine learning techniques for heart disease datasets: a survey. In: Proceedings of the 2019 11th Int Conf on Machine Learning and Computing. 2019, pp. 27–35.
[4] Lee S-H. Feature selection based on the center of gravity of BSWFMs using NEWFM. Engineering Applications of Artificial Intelligence 2015; 45: 482–487.
[5] Hu X, Pedrycz W, Wang X. Fuzzy classifiers with information granules in feature space and logic-based computing. Pattern Recognition 2018; 80: 156–167.
[6] Koulaouzidis G, Iakovidis D, Clark A. Telemonitoring predicts in advance heart failure admissions. International journal of cardiology 2016; 216: 78–84.
[7] Zadeh LA. Fuzzy logic. Computer 1988; 21: 83–93.
[8] Ojha V, Abraham A, Snášel V. Heuristic design of fuzzy inference systems: A review of three decades of research. Engineering Applications of Artificial Intelligence 2019; 85: 845–864.
[9] Blake CL, Merz CJ. UCI repository of machine learning databases, 1998.
[10] Drake J, Hamerly G. Accelerated k-means with adaptive distance bounds. In: 5th NIPS workshop on optimization for machine learning. 2012.