Adaptable Web Prediction Framework for Disease Prediction Based on the Hybrid Case Based Reasoning Model

Bruno Trstenjak
Department of Computer Engineering
Medimurje University of Applied Sciences, Cakovec, Croatia
btrstenjak@mev.hr

Dzenana Donko
Department of Computer Science
Faculty of Electrical Engineering, Sarajevo, Bosnia and Herzegovina
ddonko@etf.unsa.ba

Zikrija Avdagic
Department of Computer Science
Faculty of Electrical Engineering, Sarajevo, Bosnia and Herzegovina
zikrija.avdagic@etf.unsa.ba

***Abstract*—Nowadays, we are witnessing the rapid development of medicine and various methods that are used for early detection of diseases. In order to make quality decisions in diagnosis and prevention of disease, various decision support systems based on machine learning methods have been introduced in the medical domain. Such systems play an increasingly important role in medical practice. This paper presents a new web framework concept for disease prediction. The proposed framework is object-oriented and enables online prediction of various diseases. The framework enables online creation of different autonomous prediction models depending on the characteristics of diseases. Prediction process in the framework is based on a hybrid Case Based Reasoning classifier. The framework was evaluated on disease datasets from public repositories. Experimental evaluation shows that the proposed framework achieved high diagnosis accuracy.**

Keywords-disease prediction; web framework; hybrid model; Case Based Reasoning

I. INTRODUCTION

Data mining and prediction in medical domain is gradually gaining significant importance. In order to ensure disease prevention and quality prediction of a potential disease, support systems based on a different prediction framework have become important tools in disease diagnosis. A "framework" is the environment where a prediction process is conducted. Predicting a disease is a process of extracting hidden information from medical data and predicting the direction of disease development. For medicine purposes, many studies have been conducted with the aim to develop frameworks for disease prediction based on machine learning techniques..

Each framework has embedded one or more machine learning techniques. Today, for data prediction purposes various classification techniques are used, such as Neural networks [1], Support Vector Machines [2], Naive Bayes Classifiers [3] and many others. A framework can also be implemented by using the hybrid prediction model. The hybrid classification model consists of several machine learning methods and provides a slightly different classification approach. Techniques that are embedded in the hybrid prediction model supplement each other and contribute to the quality of prediction accuracy. A web framework is based on the hybrid CBR classification model. The hybrid model merges K-means technique for data clustering and Case Based Reasoning (CBR) classification technique.

In this paper, we propose a new concept of framework that allows online prediction and creation of different disease predictive models in a single environment. The study was conducted with the aim to develop a framework that will achieve a high degree of prediction accuracy regardless of the nature of the disease. The framework enables on-line creation of an autonomous prediction model based on disease features. The concept of the new prediction framework provides the existing Medical Decision Support Systems (MDSS) connection on the framework and starts up the prediction process based on the selected disease model. The connected MDSS, using standard web protocol, sends the collected data to the framework. Based on the request sent, the framework performs prediction and sends the obtained result of prediction to MDSS as a response to the request. In this way the adaptive web framework becomes completely independent of the medical system that uses it.

II. LITERATURE REVIEW

Implementation of different decision systems and frameworks for early disease prediction has become extremely important in the medicine domain and disease diagnosis. There is an increasing amount of research engaged in disease prediction using different machine learning techniques. With the development of machine learning techniques and hybrid models, different frameworks aiming to support disease prediction were simultaneously developed. For example, a framework based on the multi-layer classifier ensemble in the framework was presented in [4]. The proposed framework is based on the optimal combination of heterogeneous classifiers. The proposed model overcomes the limitations of conventional performance bottlenecks by utilizing an ensemble of seven heterogeneous classifiers. Evaluation shows that the proposed framework dealt with all types of attributes and achieved high diagnosis accuracy. A new framework for prediction of heart disease based on the classification techniques like Naive Bayes and Artificial Neural Networks was presented in [5]. The

authors used attribute filtering techniques like Principle Component Analysis and Information Gain for feature selection in the given data set of heart disease symptoms. In [6], a support framework for heart disease prediction using Majority Vote Based classifier ensemble was proposed. The five heterogeneous classifiers used to construct the ensemble model are as follows: Naïve Bayes, decision tree based on Gini Index, decision tree based on information gain, memory-based learner and support vector machine. Comparison of proposed framework with individual classifiers shows an increase in average accuracy. In [7], an automatic classification and prediction of Parkinson's disease was proposed. Support Vector Machine (SVM) classifier and logistic regression are used for framework construction. The framework achieved high prediction accuracy. Prediction improvements can be made by incorporating ensemble classifier instead of a single classifier.

In [8], a prediction system and a new approach for detection of heart disease based on Naive Bayes classifier were presented. This prediction system classifies medical data into five different categories and also predicts the risk of the heart disease if unknown sample is given as an input. The evaluation results confirmed a good concept of the prediction system, the classification model has achieved a high degree of prediction accuracy, the accuracy in the amount of approximately 80%. In [9], a clinical decision support system for risk prediction of heart patients was presented. The proposed system uses weighted fuzzy rules and decision tree rules for prediction. Risk prediction is performed in two phases: automated approach for generation of weighted fuzzy rules and decision tree rules; the second one is developing a fuzzy rule-based. In [10], a framework which can select the best model to predict liver disease caused by Hepatitis C Virus (HCV) was presented. The framework contains three phases: a preprocessing phase to prepare the data for applying Data Mining (DM) techniques, a DM phase to apply different DM techniques, and an evaluation phase to evaluate and compare the performance of the built models and select the best model as the recommended one. Different DM techniques had been applied: associative classification, artificial neural network, and a decision tree to evaluate the framework.

III. RESEARCH METHODOLOGY AND THE PROPOSED FRAMEWORK

The objective of our study is to develop a web prediction framework based on the hybrid CBR model for disease prediction. In this section, firstly we explain the work principle of our proposed framework. We present the structure of the proposed framework and its major components. The second part of the section describes the structure of the hybrid CBR model and used machine learning techniques.

A. Case Based Reasoning (CBR)

The hybrid prediction model is based on a Case Based Reasoning (CBR) classifier. The principle of the CBR method is based on solving new problems by observing the similarity with the previously solved problems. The CBR method uses a problem-solving approach analogous to the way of problem solving by man when he draws on his experiences. Each CBR system contains an embedded library of the cases that were resolved in the past. This is something like collecting life experiences in the domain of the problem. Each case represents a description of the problem with its associated solution. The CBR method with a built-in function of similarities tries to find the most similar case from the library. The retrieved cases from the library are used to suggest a solution. If the proposed solution is not satisfactory, the method tries to revise selected cases and find a new solution. The method adds a new revised case to the cases library and thereby expands the knowledge base. The whole execution cycle of the algorithm can be divided into four main steps: Retrieve, Reuse, Revise, Retain [11].

CBR performs measurement of similarity on the local and global level. Local similarity refers to the measurement of similarity between pairs of features. Global similarity refers to comparison of similarity between all the features that make up the object. Measuring similarity can be shown by:

$$Similarity(T,S) = \frac{\sum_{i=1}^{n} f(T_j, S_i) \times w_i}{\sum_{i=1}^{n} w_i} \qquad (1)$$

Where T=target case, S=source case, n=number of features in each case, I=individual feature from 1 to n, f=similarity function for features I in cases T and S, w=importance weighting of feature I

B. k-means

K-means is one of the simplest unsupervised learning algorithms used for solving clustering problems. Let X={xi, i=1,..,n} be a set of n dimensional objects, which should be classified into k clusters, C={cj, j=1,..,k}. The algorithm determines the quality of the clustering calculating square error between the mean of the cluster and points in the cluster. The goal of the algorithm is to minimize the sum of the squared error over all K clusters. The quality is determined by following the error function, as shown in [12]:

$$E = \sum_{j=1}^{k} \sum_{x \in C_j} \left| x_i - \mu_j \right|^2 \qquad (2)$$

where E is a sum of the squared error of all objects, μj indicates the average of cluster Cj. $|x_i-\mu_j|^2$ is a chosen distance measure between data point xi and the centroids value. The algorithm can use different methods to calculate the distance (Euclidean, Manhattan, Minkowski, etc.) [13].

C. Techniques for ranking the features

Ranking the features is a procedure which determines the importance and influence of features on the final prediction result. Due to different data types that can be used to describe a disease in the process of creating predictive model, three techniques of ranking features were used: Information Gain (IG) [14], Gain Ration (GR) [15] and Correlation-based

Feature Selection (CFS) [16]. All these techniques use entropy as a basic measure for ranking the features.

Entropy is a measure of disorderliness of the system. IG method calculates the value of the features information. The value is defined as the amount of information, provided by the feature items for the class. IG uses the following expression for the calculation:

IG (Class, Feature) = H (Class) – H (Class |Feature)

where H is entropy, which is defined by :

$$Entropy(S) = -\sum_{j=1}^{m} p_j \log_2 p_j \quad (3)$$

where p is probability, for which a particular value occurs in the sample space S. Entropy value ranges from 0 to 1. Value 0 means that all variable instances have the same value, value 1 equals the number of instances of each value. Entropy shows how the attribute values are distributed and indicates the "purity" of features.

Gain Ratio is a technique for selecting features created by extending the Information Gain technique using decision tree method. Gain ratio takes number and size of tree branches into account when choosing a feature. GR corrects the information gain by taking the intrinsic information. Intrinsic information is entropy of distribution of instances into branches (i.e. how much info we need to tell which branch the instance belongs to). Attribute value decreases as intrinsic information gets larger [17].

$$Gain\,ration\,(Attribute) = \frac{Gain\,(Attribute)}{Intrinsic\_\inf o\,(Attribute)} \quad (4)$$

CFS makes analysis and measures how strongly one attribute implies the other, based on the available data. The correlation between two variables is a goodness measure, a feature is good if it is highly correlated to the class but not highly correlated to any other feature. The evaluation measures the weight of the feature by measuring the correlation (Pearson's) between it and the class.

D. Overview of the framework structure

The framework is designed as a web environment that is accessed via the Internet. Figure 1 illustrates the framework components and their corresponding function in detail. The structure of frameworks is composed of two main components: a component for creating a model for the disease prediction according to disease characteristics, and another component which is responsible for performing the prediction on the basis of selected disease.

Creating a predictive model begins with uploading a dataset in the framework. The framework in the process of creating a predictive model uses four internal modules. All modules in Figure 1 are marked with Roman numbers: I: Data pre-processing, II: Ranking dataset features, III: Data clustering, IV: CBR classification generator. Data pre-processing includes data cleaning, normalization, transformation, feature extraction etc. This is a process of transforming raw data into a suitable format ready to be used by a data mining process. Before starting the process of creating the prediction model, the correctness of input data should be verified. The quality of prediction of the future model directly depends on the quality of input data. In this step of creating a model, the framework verifies the syntax format of the input data and determines the data type of each feature. Only in the case that data irregularity is not detected, the process of ranking the features begins.

Each disease is described by a set of properties. The input dataset, in standard CSV format, contains instances composed of the features which describe the disease. If during the preprocessing irregularity in the input dataset is not detected, the framework begins with determining the weight values for disease features. Weight values determine ranks of the features, their significance in the process of prediction. The ranking is performed using three techniques in the order to get more accurate values. After the rank module, follows a module for data clustering. The module has implemented k-means technique. To all instances from dataset, the framework determines the cluster label. The cluster instance label is determined according to similarity between instances.

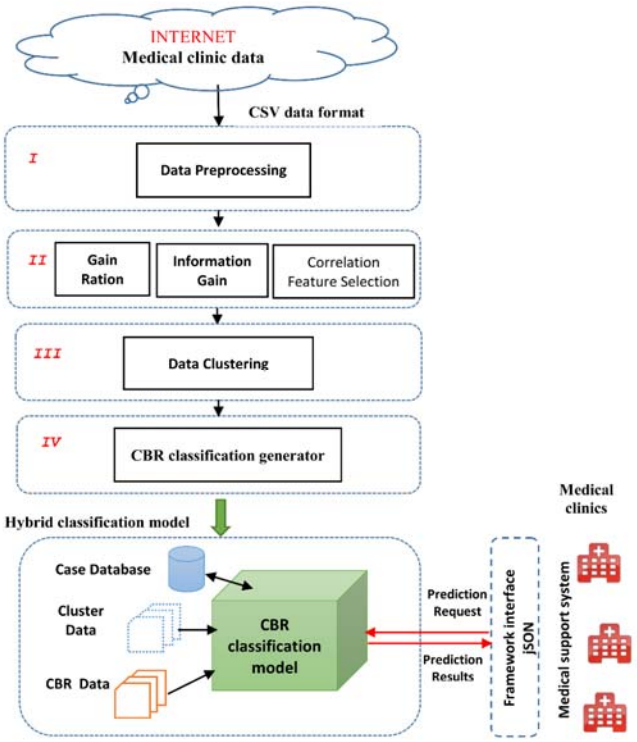

Fig. 1. The framework structure.

Last of the modules in the framework is the module responsible for the generation of a CBR classifier. This is the most complex internal process in the framework. The module creates a case database and all the elements necessary for dynamic creation of the future CBR classifier. Figure 1 shows a CBR classification model to which the case database, a cluster data component for clustering instance in the prediction process

and CBR data necessary for dynamic forming a classifier are connected. When the request for the disease prediction is sent by a user, the framework loads all these elements and adapts the CBR classifier according to the characteristic of the selected disease.

The second component in the framework is intended to perform the prediction process. Framework is not designed as a standard web portal, but as an environment to which various medical support systems can be connected. Online connection between prediction frameworks and decision system is achieved via an Internet protocol. Medical decision system sends data, features values about the disease and the request of prediction. The framework on the base of selected disease and features values, performs prediction and sends a result to the medical system. The entire communication is carried out via the specially designed framework interface.

IV. Overview of the Work Principle

The work principle of the framework is very simple. Figure 2 illustrates the work flow of the prediction process in the framework.

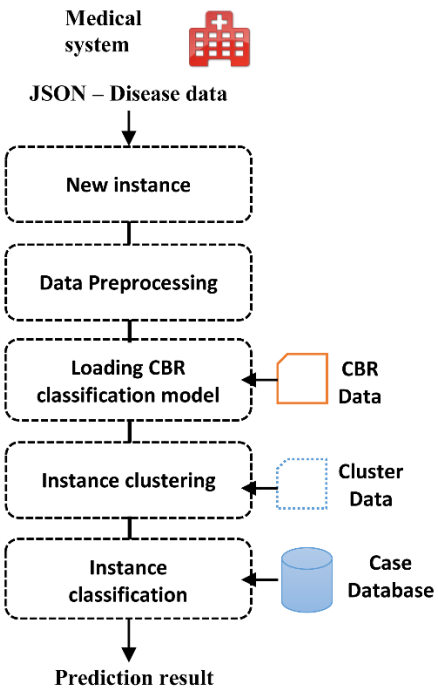

Fig. 2. Prediction process work flow.

The prediction process begins with sending a packet of information about the disease to the framework. The user through the medical decision system enters data about the disease. Medical decision system sends data to the framework in JSON data format. The framework creates received data into a new instance whose structure is adapted to the structure of case database of the selected disease. A newly formed instance is forwarded to a pre-processing process to verify the correctness of the features value. If irregularity in the instance structure is not detected, the framework begins the adaptation of the hybrid model based on the selected disease. The hybrid model loads the data necessary to form a CBR classifier, begins with initialization of the classification model and sets up the data about features weight values, mode of conducting the prediction, information about the structure of case database, etc. When the adaptation is completed, the framework starts with the clustering process of a new instance.

Based on the information about cluster centroids (Cluster Data), defined during the creation of the prediction model, the framework performs clustering of a new instance. The instance with additional information about the cluster label shall be forwarded to the classification. The CBR classification model, using the case database, performed the classification. The achieved result represents the result of prediction. The framework sends back prediction result to the medical support system by which the request was sent.

V. Experimental Results and Discussion

A. Dataset description

In the development of the hybrid model and evaluation of the prediction quality, a large number of data sets was used. The majority of the experiments are performed on 15 well known medical data sets publicly available from the UCI machine learning repository [18]. The decision to use these dataset is based on the tendency that the results of evaluation compare with the results achieved with a similar hybrid models. Table I shows a list of used data sets in evaluation of prediction accuracy over different hybrid models. The following data sets were used: D1: breast cancer dataset retrieved from the University Medical Centre, Institute of Oncology, Ljubljana, D2: breast cancer dataset retrieved from Wisconsin clinical sciences centre, D3: diabetes datasets are named as Pima Indian Diabetes Dataset, D4: hepatitis disease dataset.

B. Evaluation of prediction accuracy

The first part of the case study was focused on measuring the prediction accuracy of the framework. Table I shows the achieved prediction results, where CBR is prediction accuracy, PR is classification precision and SE is sensitivity achieved with the framework. During the evaluation the framework achieved a high degree of prediction accuracy and a high degree of sensitivity.

TABLE I. OVERVIEW ACHIEVED PREDICTION ACCURACY RESULTS.

No	Dataset	CBR	PR	SE
1	D1	87.71	0.76	0.78
2	D2	100.00	1.00	1.00
3	D3	98.20	0.96	0.96
4	D4	91.33	0.85	0.90

In order to obtain complete details about the characteristics of the framework, our achieved results were compared with the results achieved by the following hybrid models: DTiGP - Functional Tree (FT) classifier [19], C4.5+NB - Decision Tree + Naive Bayes classifier [20], GDADT - Genetic based Data Adaptation (GDA)+ Decision Tree (DT) [21], BFT - Best First Tree [22], REP - Regression Tree + Information Gain [22], GA+C4.5 - Genetic Algorithm(GA) + Decision Tree (RGDT)

[22], SRLPSO + ELM - self-regulated learning capability of the particle swarm optimization (PSO) algorithm with the extreme learning machine (ELM) classifier [23], HMV - Hierarchical Majority Voting [4].

TABLE II. COMPARISON RESULTS OVERVIEW.

Model	D1	D2	D3	D4
Hybrid CBR	*87.71*	*100*	*98.20*	*91.33*
DTiGP	71.70	94.70	72.65	83.20
C4.5+NB	71.66	95.77	75.79	80.98
GDADT		92.60	85.30	
HMV	85.00	97.00	77.08	86.45
BFT	69.58		72.65	80.64
REP	66.78		70.31	78.06
GA+C4.5	75.87		74.21	85.16
SRLPSO+ELM	91.33	99.78	93.09	98.71

Based on all achieved results noted in Table II, it may be seen that the framework with a CBR hybrid model achieved very good results compared to other hybrid models. With all used data sets the hybrid model achieved a high degree of prediction accuracy. Analysing the results of other hybrid models, only a SRLPSO+ELM model achieved similar results. The achieved results confirm the good concept of framework and implemented the prediction algorithm.

VI. CONCLUSION

This paper has introduced a new approach for prediction framework. A new framework merges several machine learning methods in the CBR hybrid model. The results achieved by our experiments suggest that the proposed framework possesses good properties from the standpoint of quality. The concept of the framework provides the dynamic adaptation, depending on the disease characteristics. This research will be followed by additional testing of the framework in online environment.

REFERENCES

[1] F. Amato, A. Lopez, E. M. Pena-Mendez, P. Vanhara, A. Hampl, J. Havel, "Artificial neural networks in medical diagnosis", Journal of Applied Biomedicine, Vol. 11, pp. 47-58, 2013

[2] C. C. Sady, A. L. P. Ribeiro, "Symbolic features and classification via support vector machine for predicting death in patients with Chagas disease", Computers in Biology and Medicine, Vol. 70, pp. 220-227, 2016

[3] M. -C. Yang, C. -S. Huang, J. -H. Chen, R.-F. Chang, "Whole Breast Lesion Detection Using Naive Bayes Classifier for Portable Ultrasound", Ultrasound in Medicine & Biology, Vol. 38, No. 11, pp. 1870–1880, 2012

[4] S. Bashir, U. Qamar, F. H. Khan, L. Naseem, "HMV: A medical decision support framework using multi-layerclassifiers for disease prediction", Journal of Computational Science, Vol. 13, pp. 10-25, 2016

[5] K. Sudhakar, D. M. Manimekalai, "Propose a Enhanced Framework for Prediction of Heart Disease", Int. Journal of Engineering Research and Applications, Vol. 5, No. 4, pp. 1-6, 2015

[6] S. Bashir, U. Qamar, F. H. Khan, M. Y. Javed, "MV5: A Clinical Decision Support Framework for Heart Disease Prediction Using Majority Vote Based Classifier Ensemble", Arabian Journal for Science and Engineering, Vol. 39, No. 11, pp. 7771-7783, 2014

[7] R. Prashantha, S. D. Roya, P. K. Mandal, S. Ghosh, "Automatic classification and prediction models for early Parkinson's disease diagnosis from SPECT imaging", Expert Systems with Applications, Vol. 41, No. 7, pp. 3333-3342, 2014.

[8] D. S. Medhekar, M. P. Bote , S. D. Deshmukh, "Heart Disease Prediction System using Naive Bayes", International Journal of Enhaced Research in Science Technology & Engineering, Vol. 2, No. 3, pp. 1-5, 2013

[9] P. K. N. Anooj, "Clinical decision support system: risk level prediction of heart disease using weighted fuzzy rules and decision tree rules", Central European Journal of Computer Science, Vol. 1, No. 4, pp. 452-498, 2011

[10] E. M. F. El Houby, "A Framework for Prediction of Response to HCV Therapy Using Different Data Mining Techniques", Advances in Bioinformatics, Vol. 2014, Article ID 181056, pp. 1-10, 2014

[11] M. M. Richter, R. Weber, "Case-Based Reasoning: A Textbook", in Basic CBR Elements, Springer Science & Business Media, pp. 17-34, 2013

[12] A. Jain, "Data clustering: 50 years beyond K-means", Pattern Recognition Letters, Vol. 31, No. 8, p. 651–666, 2010

[13] V. R. Patel, R. G. Mehta, "Data Clustering: Integrating Different Distance Measures with Modified k-Means Algorithm", Advances in Intelligent and Soft Computing, Vol. 131, pp. 691-700, 2012

[14] B. Azhagusundari, A. S. Thanamani, "Feature Selection based on Information Gain", International Journal of Innovative Technology and Exploring Engineering, Vol. 2, No. 2, pp. 18-21, 2013

[15] F. Li, Y. Piao, M. Li, M. Piao, K. Ryu, "Positive impression of low-ranking microrn as in human cancer classification", in 4th International conference on Computer Science & Information Technology, Sydney, Australia, 2014

[16] R. Priyadarsini, M. Valarmathi, S. Sivakumari, "Gain Ration based feature selection method for privacy preservation", ICTACT Journal on Soft Computing, Vol. 1, No. 4, pp. 201-205, 2011

[17] R. Tiwari, M. P. Singh, "Correlation-based Attribute Selection using Genetic Algorithm", International Journal of Computer Applications, Vol. 4, No. 8, pp. 28-34, 2010

[18] C. L. Blake, C. J. Merz, UCI Repository of machine learning, University of California, Department of Information and Computer Science, 1998

[19] R. Konig, U. Johansson, T. Lofstrom, L. Niklasson, "Improving GP Classification Performance by Injection of Decision Trees", WCCI 2010 IEEE World Congress on Computational Intelligence, Barcelona, Spain, 2010

[20] L. Jiang, C. Li, "Scaling Up the Accuracy of Decision-Tree Classifiers: A Naive-Bayes Combination", Journal of Computers, Vol. 6, No. 7, pp. 1325-1331, 2011

[21] R. G. Ramani, L. Balasubramanian, A. A. Meenal, "A hybrid classification model employing Genetic algorithm and Root Guided Decision Tree for improved categorization of data", ARPN Journal of Engineering and Applied Sciences, Vol. 10, No. 21, pp. 9968-9975, 2015

[22] P. Lakshmi, S. S. Kumar, A. Suresh, "A Novel Hybrid Medical Diagnosis System Based on Genetic Data Adaptation Decision Tree and Clustering", ARPN Journal of Engineering and Applied Sciences, Vol. 10, No. 16, pp. 7293-7299, 2015

[23] C. V. Subbulakshmi, S. N. Deepa, "Medical Dataset Classification: A Machine Learning Paradigm Integrating Particle Swarm Optimization with Extreme Learning Machine Classifier", The Scientific World Journal, Vol. 2015, Article ID 418060, pp. 1-12, 2015