ISSN 1870-4069

Computational Intelligence Algorithms Applied to the Pre-diagnosis of Chronic Diseases

Mariana Dayanara Alanis-Tamez[1], Yenny Villuendas-Rey[2], Cornelio Yáñez-Márquez[1]

[1] Instituto Politécnico Nacional, Centro de investigación en Computación, Mexico

[2] Instituto Politécnico Nacional, Centro de Innovación y Desarrollo Tecnológico en Cómputo, Mexico

Abstract. Classification models applied to medicine have become an increasing area of research worldwide. Such as, the application and development of known models and algorithms for disease diagnosis and prediction have been an active research topic. The present article is a study of the classification algorithms most used in the literature, and its application to the diagnosis of chronic diseases. More specifically, we tested five classification models, over medical data. The application of the supervised classification algorithms is done over the Knowledge Extraction based on Evolutionary Learning (KEEL) environment, using a Distributed optimally balanced stratified 5-fold cross validation scheme. In addition, the experimental results obtained were validated to identify significant differences in performance by mean of a non-parametric statistical test (the Friedman test). The hypothesis testing analysis of the experimental results indicates which supervised classification model outperforms others for medical diagnosis.

Keywords: classification models, medical informatics, disease prediction and diagnosis, computational intelligence.

1 Introduction

For medical diagnosis, computational applications have been developed mainly in two principal areas: algorithm development and system development for supporting diagnosis. Among others, several web-based systems, as well as tools for telediagnosis and telemedical devices can be mentioned [1, 2, 3].

Chang et al. [1] developed a Web-based decision support system that considers the sensitivity analysis as well as the optimal prior and posterior decisions applied for some chronic diseases. The purpose of their work is to review several approaches and to develop a Web-based decision support system (DSS).

Sanchez-Santana et al. [2] introduced another computational approach for medical diagnosis. They propose a new telediagnosis environment for the detection of cardiovascular problems and allows medical staff to semi-automatically identify and quantify a patient's potential cardiovascular complications.

In addition, Havlik et al [3] propose a solution for remote monitoring, rapidly developing devices for telemedical applications and assistive technologies. The approach used was to design and realize a modular system consisting of input modules for signal acquisition, a control unit for signal pre-processing, handshaking of data communication, controlling the system and providing the user interface and communication modules for data transmission to a superordinate system.

In this paper, for assisting medical diagnosis, several algorithmic solutions have been proposed in recent years (section 2). We study the different classification algorithms most used in the literature, and its application to the diagnosis of chronic diseases (section 3). The different datasets have: imbalanced data, mixed categorical and numerical attributes, and missing values. In addition, we used 15 medical related datasets (section 4) to test all the classification models and the numerical experiments performed in the comparison between the different classification algorithms applied to the medical diagnosis. Finally, we offer the conclusion and some lines of future work (section 5).

2 Background

For medical diagnosis, several algorithmic solutions have been proposed in recent years.

Aiping Lu et al. [4] used pattern classification as a guideline in disease classification in Traditional Chinese medicine (TCM) practice and has been recently incorporated with biomedical diagnosis. They describe the historical evolution on the integration of the TCM pattern classification and disease diagnosis in biomedicine, the methodology of pattern classification for diseases

Timothy J. W. Dawes et al. [5] used a machine-learning survival model that uses three-dimensional cardiac motion which predicts outcome independent of conventional risk factors in patients with newly diagnosed pulmonary hypertension. It is used to determine if patient survival and mechanisms of right ventricular failure in pulmonary hypertension could be predicted by using supervised machine learning of three-dimensional patterns of systolic cardiac motion.

Yu Sun et al. [6] used the performance of a support vector machines (SVM) algorithm to predict prostate tumor location using multi-parametric MRI (mpMRI). They used a Gaussian kernel SVM which was trained and tested on different patient data subsets. Parameters were optimized using leave-one out cross validation.

Fazekas [7] addressed the examination of the periodicity of the childhood leukemia in Hungary using seasonal decomposition time series. The dataset used was from the Hungarian Pediatric Oncology Workgroup, and contained the data of all the patients with lymphoid leukemia diagnosed between 1988 and 2000. These data highlight the role of the environmental effects, like viral infections, epidemics, among others on the onset of the disease.

To sum up, Zheng et al. proposed a data informed framework for identifying subjects with and without Type 2 Diabetes Mellitus from Electronic Health Records via feature engineering and machine learning [8]. The objective of this work is to develop

a semi-automated framework based on machine learning as a pilot study to liberalize filtering criteria to improve recall rate with a keeping of low false positive rate.

3 Datasets and Algorithms

In this section, a summary of the classification algorithms applied and the datasets related to medical diseases are presented.

3.1 Datasets Related to Medical Diseases

The used datasets include information about different diseases, such as breast cancer, thyroid diseases, heart diseases, liver disorders, diabetes, and others, all related to mainly with chronic diseases. The datasets used in this paper were taken from online information provided by KEEL dataset repository [9]. We used the medical related standard classification datasets.

In the following, a description of each of the datasets used is shown.

Breast Cancer dataset: It is one of three domains provided by the Oncology Institute, which has repeatedly appeared in the machine learning literature. This dataset has two classes, 201 instances of one class and 85 instances of another class. The instances are described by nine attributes, some of which are linear and some are nominal.

Liver Disorders (BUPA) dataset: It analyzes some liver disorders that might because of the excessive alcohol consumption. The objective is to select if a given individual suffers from alcoholism.

Heart Disease (Cleveland) dataset: It is a part of the Heart Disease Dataset (the part obtained from the V.A. Medical Center, Long Beach and Cleveland Clinic Foundation), using a subset of 14 attributes. The objective is to detect the presence of heart disease in the patient.

Haberman's Survival dataset: It contains cases from a study that was conducted between 1958 and 1970 at the University of Chicago's Billings Hospital on the survival of patients who had undergone surgery for breast cancer. The objective is to determine if the patient survived 5 years or longer (positive) or if the patient died within 5 years (negative).

Statlog (Heart) dataset: The objective is to detect the absence or presence of heart disease.

Hepatitis dataset: It contains a mixture of integer and real valued attributes, with information about patients affected by the hepatitis disease. The objective is to predict if these patients will die or survive.

Mammographic Mass dataset: It can be used to predict the severity (benign or malignant) of a mammographic mass lesion from BI-RADS attributes and the patient's age. The data was collected at the Institute of Radiology of the University Erlangen-Nuremberg between 2003 and 2006.

Thyroid Disease (New Thyroid) dataset: It is one of the several databases about Thyroid available at the UCI repository. The objective is to detect is a given patient is normal or suffers from hyperthyroidism or hypothyroidism.

Pima Indians Diabetes dataset: The diagnostic, binary-valued variable investigated is whether the patient shows signs of diabetes according to World Health Organization criteria

Post-Operative dataset: The classification objective of this database is to determine where patients in a postoperative recovery area should be sent to next. Because hypothermia is a significant concern after surgery, the attributes correspond roughly to body temperature measurements.

South African Hearth dataset: A retrospective sample of males in a heart-disease high-risk region of the Western Cape, South Africa. The class label indicates if the person has a coronary heart disease: negative or positive.

SPECTF Heart dataset: It describes diagnosing of cardiac Single Proton Emission Computed Tomography (SPECT) images. Each of the patients is classified into two categories: normal or abnormal.

Thyroid Disease (thyroid0387) dataset: The objective is to detect is a given patient is normal or suffers from hyperthyroidism or hypothyroidism. This dataset is one of the several databases about Thyroid available at the UCI repository.

Breast Cancer Wisconsin (Diagnostic) dataset: It contains 30 features computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. The objective is to determine if a found tumor is benign or malignant.

Breast Cancer Wisconsin (Original) dataset: It contains cases from a study that was conducted at the University of Wisconsin Hospitals, Madison, about patients who had undergone surgery for breast cancer. The objective is to determine if the detected tumor is benign or malignant.

Table 1. Description of the datasets used.

	Datasets	Attributes		Imbalance analysis		Missing values	Classes
		Numeric	Categorical	Instances	IR		
1.	breast	0	9	277	2.420	Yes	2
2.	bupa	7	0	345	1.379	No	2
3.	cleveland	13	0	297	12.308	Yes	5
4.	haberman	3	0	306	2.778	No	2
5.	heart	13	0	270	1.250	No	2
6.	hepatitis	19	0	80	5.154	Yes	2
7.	mammographic	6	0	830	1.060	Yes	2
8.	newthyroid	5	0	215	5.000	No	3
9.	pima	8	0	768	1.866	No	2
10.	post-operative	0	8	87	62.000	Yes	3
11.	saheart	8	1	462	1.888	No	2
12.	spectfheart	44	0	267	3.855	No	2
13.	thyroid	21	0	7200	40.157	No	3
14.	wdbc	30	0	569	1.684	No	2
15.	wisconsin	9	0	683	1.858	Yes	2

In table 1, a summary of the description of the datasets is given. The summary includes the amount of numerical and categorical attributes, the number of instances, the Imbalance Ratio (IR) among majority and minority classes, the presence or not of missing values, and the number of classes.

For validation purposes, we used the Distributed optimally balanced stratified cross validation procedure (dob-scv) with five folds, introduced by [10] and recommended for imbalanced scenarios.

3.2 Performance Measures and Statistical Tests

Firstly, when imbalanced datasets were used for classification, the usual performance measures become inadequate [11]. This is because of the bias that such measures have towards the majority class, which in turn may yield to misleading conclusions. For evaluating the performance over imbalanced datasets with multiple classes, the use of the average True Positive Rate for each class [11] have been raised.

Therefore, in a two classes problem, the true positive rate (TPR) (as well-known as recall or sensitivity) considers the total of positive instances correctly classified, relative to the total of instances of the positive class, considering True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). See equation (1).

$$Classification\ Sensitivity = TPR = Recall = \frac{TP}{TP + FN}. \tag{1}$$

In fact, in a problem with k classes, the classification sensitivity takes into consideration the total of correctly classified instances from class j, relative to the total of instances of the j-th class. Therefore, the classification sensitivity for class j calculate the probability of correctly classifying an instance from class j. For the computation of such classification sensitivity, let n_j be the number of correctly classified instances (in a confusion matrix of k classes), and let t_j be the total of instances belonging to class j. So, for this reason the classification sensitivity (also recall or true positive rate) of class j, denoted by S_j, is computed as follows:

$$S_j = Recall_j = TPR_j = \frac{n_j}{t_j}. \tag{2}$$

The minimum classification sensitivity is given by [7]:

$$Minimum = \min_{j=1..k}\{S_j\}. \tag{3}$$

Although minimum classification sensitivity allows handling multiple classes, it only considers the lesser of the correctly classified rates among the classes. That is why in this paper, a measurement of performance is giving the same weight to each of the classes, independently from the number of samples each has, was chosen. The performance measure used here is the average classification sensitivity per class [10] which is defined as:

$$Average = \frac{1}{k}\sum_{i=1}^{k} S_j. \tag{4}$$

In the previous equation k is the number of classes and S_j is the classification sensitivity for the j-th class. This performance measure allows us to evaluate the global performance of classification algorithms over all the classes in the problem, not only

over the minority class. The use of the average classification sensitivity per class allows taking into consideration all the classes, without bias towards any one. Figure 3 presents an example of how the minimum classification sensitivity and the average classification sensitivity are computed, with k = 3 classes:

		Real Class		
		X	***Y***	***Z***
Predicted class	***X***	2	8	0
	Y	3	3	4
	Z	6	2	2

$$S_A = \frac{2}{10} = 0.2,\ S_B = \frac{3}{10} = 0.3,\ S_C = \frac{2}{10} = 0.2$$

a) $Average = \frac{0.2+0.3+0.2}{3} = \frac{0.7}{3}$ b) $Minimum = min\{0.2, 0.3, 0.2\} = 0.2$

Fig. 1. An example of computation of performance measures given a confusion matrix for three classes taking in consideration the average classification sensitivity and the minimum classification sensitivity.

To determine which classification algorithms got the better experimental results while predicting voting intentions and hypothesis testing was used. Statistical hypothesis tests evaluate whether there is a significant difference in the performance given by different classification algorithms, talking about their prediction of the voting intentions. Regarding the works of [12,13], non-parametric tests were chosen for the current research. Particularly, the Friedman test is widely recommended for this kind of works so it was selected.

In addition, the Friedman test is a statistical non-parametric test developed by Friedman [14], which turns out to be the non-parametric equivalent to the two-way ANOVA analysis. The Friedman test consists of ordering the samples and replacing them by their respective ranks like: the best result corresponds to rank 1, the second best to rank 2, the third to rank 3 and so on. After that, the existence of identical samples is taken into consideration, in which case they are assigned an averaged rank.

3.3 Classification Algorithms

Firstly, a decision tree [15] is a predictor that indicates the label associated with an instance by traveling from a root node of a tree to a leaf, the leaf will be the assigned class. At each node on the root to leaf, the successor child in the tree is chosen based on a splitting of the input space. **C4.5** is a decision tree which is an extension of the ID3 algorithm, this algorithm is almost always referred to as a statistical classifier.

Another algorithm used is **kNN** [15] or *k* nearest neighbor algorithm, it is a supervised classification method whose training phase consists of storing the characteristic vectors and class labels of training examples. The distance between the stored vectors is calculated and the "*k*" nearest instances are selected; the selected instance is sorted with the most repeating class.

Logistic [15] or Logistic Regression, is a type of regression analysis used to predict the outcome of a categorical variable as a function of independent or predictor variables. However, logistic regression is used for classification tasks: We can interpret *h(x)* as the probability that the label of *x* is 1.

MLP [15] or Multilayer Perceptron is a network of simple "neurons" or "perceptrons". It has its neurons grouped in layers of distinct levels. Each of the layers is formed by a set of neurons and three different types of layers are distinguished: the input layer, the hidden layers and the output layer.

Finally, **SMO** [16] or Sequential minimal optimization is an algorithm for solving the quadratic programming problem that arises during the training of support vector machines, it was implemented by John Platt's for the machines learning tool called WEKA (Waikato Environment for Knowledge Analysis).

The classification algorithms described in this section were used to evaluate the datasets proposed in the next section.

4 Experimental Results and Discussion

This section presents the experimental results obtained for medical diagnosis, using the classification models proposed, as well as other state of art models. Figure 2 illustrates the schematics of the experiment design carried out.

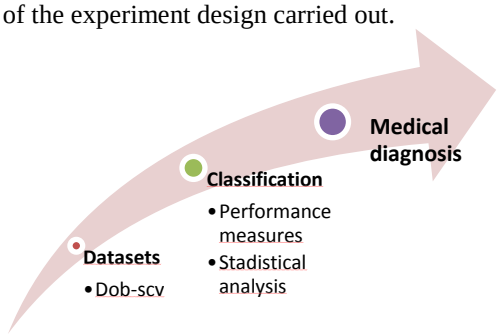

Fig. 2. Schematic of the experimental design.

Five state-of-the-art classification algorithms were selected. All of them can deal with mixed and incomplete data. This selection includes Nearest Neighbor (NN) (using k=3), Multilayer Perceptron (MLP), C4.5, SMO and Logistic. For MLP, C4.5, Logistics and SMO, we used the default parameter values offered in the KEEL software package.

For the Nearest Neighbor classifier, the HEOM dissimilarity was used [17] which handles mixed and incomplete data descriptions. Considering that max_a and min_a are the maximum and minimum values of the feature a, if it is a numerical attribute, the HEOM dissimilarity is computed as:

$$HEOM(x,y) = \sqrt{\sum_{a=1}^{m} d_a(x_a, y_a)}. \quad (5)$$

$$d_a \begin{cases} 1 \text{ if } x_a \text{ or } y_a \text{ are unknown,} \\ \text{overlap } (x_a, y_a) \text{ if } a \text{ is numerical,} \\ diff(x_a, y_a) \text{ if a is categorical,} \end{cases}$$

$$overlap\ (x_a, y_a) = \begin{cases} 0 \text{ if } x_a = y_a, \\ 1 \text{ otherwise,} \end{cases}$$

$$diff\ (x_a, y_a) = |x_a - y_a| / / \ max_a - min_a\ .$$

Table 2 shows the results obtained by the analyzed classification algorithms, for the medical diagnosis problems considered. Best results are highlighted in bold.

Table 2. Average True Positive Rate obtained by the classification algorithms.

Datasets	C4.5	kNN	Logistic	MLP	SMO
breast	0.591	0.605	0.595	**0.659**	0.632
bupa	0.614	0.652	**0.659**	0.535	0.500
cleveland	0.292	0.297	**0.319**	0.298	0.310
haberman	0.578	0.583	0.564	**0.649**	0.500
heart	0.775	0.803	**0.835**	0.833	0.833
hepatitis	0.679	0.732	0.641	**0.820**	0.693
mammographic	**0.838**	0.818	0.828	0.459	0.824
newthyroid	0.894	0.914	**0.956**	0.695	0.767
pima	0.687	0.690	**0.730**	0.708	0.714
post-operative	0.328	0.343	0.326	**0.641**	0.336
saheart	0.618	0.607	0.669	0.643	**0.688**
spectfheart	0.565	**0.701**	0.606	0.579	0.509
thyroid	**0.976**	0.593	0.724	0.447	0.518
wdbc	0.479	0.475	0.487	**0.500**	0.477
wisconsin	0.502	0.511	**0.512**	0.510	0.503
Times Best	2	1	**6**	5	1

To find the best classification algorithm is more appropriate for the correct diagnosis of diseases, the Friedman test [13] was applied.

Table 3. Algorithms rankings according to the Friedman: the best performer is Logistic.

No.	Algorithm	Ranking
1	Logistic	2.267
2	MLP	2.767
3	kNN	3.000
4	SMO	3.300
5	C4.5	3.667

The algorithms rankings according to the Friedman test are shown in table 3, where the best classifier for this task is, apparently, Logistic.

Considering the results of the Friedman Test, it rejects the hypothesis having an adjusted value lower or equal than 0.05. The test indicates that the probability of Friedman is 0.149, this value is greater than 0.05 thus the result confirms that although the Logistic model was the number one in the ranking, the Friedman test did not find significant differences in the performance of the five classifiers analyzed.

5 Conclusions and Future Work

In this paper, it is proposed to study and evaluate the different classification algorithms most used in the literature, in order to determine which of these is the best in terms of medical diagnosis. These classification algorithms were selected because all of them can deal with mixed and incomplete data.

The result was that among the selected algorithms, in the analysis of diagnosis of several diseases, the logistic model obtained very good results; due to it significantly outperform other classifiers in medical diagnosis. According to the Friedman test, the best classifier in the experiments carried out is Logistic; but, considering the results of the probability of Friedman that is 0.149 and this value is greater than 0.05 we conclude that although the Logistic model was the number one in the ranking, the Friedman test did not find significant differences in the performance of the five classifiers analyzed.

As a future work, we would propose a new algorithm, different from those already evaluated in the literature, that competes with the evaluated classifiers and that also has significant differences against the others. This new algorithm should be proposed taking into consideration that it is able of handling mixed, incomplete and multiclass data.

Acknowledgements. The authors would like to thank the Instituto Politécnico Nacional (Secretaría Académica, COFAA, SIP, CIDETEC, and CIC), the CONACyT, and SNI for their economic support to develop this work.

References

1. Chang, C. C., Cheng, C. S., Huang, Y. S.: A Web-Based Decision Support System for Chronic Diseases. JUCS (Journal for Universal Computer Science), 12(1), 115–125 (2006)
2. Sanchez-Santana, M. A., Aupet, J. B., Betbeder, M. L., Lapayre, J. C., Camarena-Ibarrola, A.: A tool for telediagnosis of cardiovascular diseases in a collaborative and adaptive approach; JUCS (Journal for Universal Computer Science), 19(9), 1275–1294 (2013)
3. Havlik, J., Lhotska, L., Parak, J., Dvorak, J., Horcik, Z., Pokorny, M.: A Modular System for Rapid Development of Telemedical Devices; JUCS (Journal for Universal Computer Science), 19(9), 1242–1256 (2013)
4. Lu, A., Jiang, M., Zhang, C., Chan, K.: An integrative approach of linking traditional Chinese medicine pattern classification and biomedicine diagnosis. Journal of ethnopharmacology, 141(2), 549–556 (2012)
5. Dawes, T. J., de Marvao, A., Shi, W., Fletcher, T., Watson, G. M., Wharton, J., Cook, S. A.: Machine learning of three-dimensional right ventricular motion enables outcome prediction in pulmonary hypertension: a cardiac MR imaging study. Radiology, 283(2), 381–390 (2017)
6. Sun, Y., Reynolds, H., Wraith, D., Williams, S., Finnegan, M. E., Mitchell, C., Haworth, A.: Predicting prostate tumor location from multiparametric MRI using Gaussian kernel

support vector machines: a preliminary study. Australasian physical & engineering sciences in medicine, 40(1), 39–49 (2017)
7. Fazekas, M.: Analysing Data of Childhood Acute Lymphoid Leukaemia by Seasonal Time Series Methods. JUCS (Journal for Universal Computer Science), 12(9), 1190–1195 (2006)
8. Zheng, T., Xie, W., Xu, L., He, X., Zhang, Y., You, M., Chen, Y.: A machine learning-based framework to identify type 2 diabetes through electronic health records. International Journal of Medical Informatics, 97, 120–127 (2017)
9. KEEL dataset repository Homepage, https://www.keel.es
10. López, V., Fernández, A., García, S., Palade, V., Herrera, F.: An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. Information Sciences, 250, 113–141 (2013)
11. Fernández, A., López, V., Galar, M., Del Jesus, M. J., Herrera, F.: Analysing the classification of imbalanced data-sets with multiple classes: Binarization techniques and ad-hoc approaches. Knowledge-based systems, 42, 97–110 (2013)
12. Demšar, J.: Statistical comparisons of classifiers over multiple datasets. Journal of Machine Learning Research, 7, 1–30 (2006)
13. García, S., Fernández, A., Luengo, J., Herrera, F.: Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power. Information Sciences, 180(10), 2044–2064 (2010)
14. Friedman, M.: A comparison of alternative tests of significance for the problem of m rankings. Annals of Mathematical Statistics, 11, 86–92 (1940)
15. Duda, R. O., Hart, P. E., Stork, D. G.: Pattern classification. Wiley, New York (1973)
16. SMO Homepage, http://weka.sourceforge.net/doc.dev/weka/classifiers/functions/SMO.html
17. Wilson, D. R., Martinez, T. R.: Improved heterogeneous distance functions. Journal of Artificial Intelligent Research, 1–34 (1997)