Journal of Universal Computer Science, vol. 24, no. 6 (2018), 775-796
submitted: 3/7/17, accepted: 23/4/18, appeared: 28/6/18 © J.UCS

Medical Diagnosis of Chronic Diseases Based on a Novel Computational Intelligence Algorithm

Yenny Villuendas-Rey
(Instituto Politécnico Nacional, CIDETEC-IPN, Ciudad de México, México
yenny.villuendas@gmail.com)

Mariana-D. Alanis-Tamez
(Instituto Politécnico Nacional, CIC-IPN, Ciudad de México, México
malanis93@hotmail.com)

Carmen-F. Rey Benguría
(Centro de Estudios "José Martí", Universidad de Ciego de Ávila, Ciego de Ávila, Cuba
carmenrb2008@gmail.com)

Cornelio Yáñez-Márquez
(Instituto Politécnico Nacional, CIC-IPN, Ciudad de México, México
cyanez@cic.ipn.mx)

Oscar Camacho-Nieto
(Instituto Politécnico Nacional, CIDETEC-IPN, Ciudad de México, México
oscarc@cic.ipn.mx)

Abstract: Computational Intelligence techniques in medicine have become an increasing area of research worldwide. Among them, the application and development of new models and algorithms for disease diagnosis and prediction have been an active research topic. The research contribution of the current paper is the proposal of a novel classification model, and its application to the diagnosis of chronic diseases. One of the main characteristics of the new model is that it is designed to deal with imbalanced data. With the purpose of making experimental comparisons to demonstrate the benefits of the proposed model, 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 in order to identify significant differences in performance by mean of a non-parametric statistical test (the Friedman test), and a post-hoc test (the Holm test). The hypothesis testing analysis of the experimental results indicates that the proposed model outperforms other supervised classifiers for medical diagnosis.

Keywords: medical informatics, disease prediction and diagnosis, computational intelligence
Categories: J.3, I.2.1, I.2.6, I.5.1

1 Introduction

The social impact of chronic diseases in the population is one of the current topics in scientific research worldwide [Fazekas, 06]. The efforts made by the research groups are remarkable, with obvious interest to minimize the negative effects of this type of maladies. For professionals involved in public health issues, it is evident that early diagnosis of a chronic disease in a patient increases their chances of survival [Abdar et al, 17].

In this context, Computational Intelligence techniques in medicine have become an increasing area of research worldwide, and the application and development of new models and algorithms for disease diagnosis and prediction is an active research topic. A careful analysis of these techniques, models and algorithms, allows us to realise that they have weak points and failures that are worth facing.

For example, some models depend on the availability of a certain type of actual failure data [Chang et al., 06]; others are modular and have the disadvantage that if any of the modules fail, the diagnosis will not be carried out [Havlik et al., 13]; some more behave like boxes, which makes impossible to determine which instances were classified incorrectly and why [Rijo et al, 14]; some, although useful in show correct early diagnoses, require pre-processing data, modifying thus the original dataset [Abdar et al, 17].

In the effort to try to minimize the disadvantages described previously, in this paper we introduce a novel classification model, designed for the diagnosis of chronic diseases. The new model diminishes the undesired effects of the curse of dimensionality; it deals with the presence of small disjoints by detecting subclasses in classes; it deals with imbalanced data, and handles mixed as well as incomplete data; besides, it deals with class overlapping. In addition, the proposed model is interpretable and transparent, due to we know exactly why an instance belong to a certain class. This is a clear advantage over other classifiers of the state of art, including the ones mentioned above.

The rest of the article is organized as follows: section 2 briefly describes several algorithmic solutions for assisting medical diagnosis, which have been proposed in recent years; Its benefits and its main disadvantages are specified. In section 3, the proposed model and its training and classification phases are explained in detail. We used several medical related datasets to test our proposal, and the numerical experiments carried out are shown in section 4. The paper ends with the conclusions and future works at Section 5.

2 Related works

In the current specialized literature several computational algorithmic solutions for medical diagnosis are reported. The most outstanding are mentioned below.

Chang et al. developed a Web-based decision support system considering the sensitivity analysis as well as the optimal prior and posterior decisions for some chronic diseases. It can deal with uncertain prior knowledge about the physiological system by considering the optimal prior decision and with the sensitivity analysis and the optimal posterior decision [Chang et al., 06]. However, it can do it only if actual failure data were available.

In addition, Havlik et al. propose a solution for rapidly developing devices for telemedical applications, remote monitoring and assistive technologies [Havlik et al., 13]. The approach was to develop 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; however, if any of the modules fail, the diagnosis will not be carried out.

Rijo et al. used a text mining approach to support medical decisions relating to epilepsy diagnosis and ICD-9-based classification in children [Rijo et al, 14]. They put forward a text mining approach using electronically processed medical records, and apply the K Nearest Neighbour technique as a white-box multiclass classifier approach to classify each instance, mapping it to the corresponding ICD-9-based standard code. Due to this classifier is treated as a white-box, it is not possible to determine which instances were classified incorrectly and why; besides that, the author proposes as future work foreseen in substantially enlarging the dataset, since having more records with other seizure type's classification, and providing the extraction of more relevant features will provide greater performance to the results.

Fazekas addressed the examination of the periodicity of the childhood leukaemia in Hungary using seasonal decomposition time series [Fazekas, 06]. The dataset used was from the Hungarian Paediatric Oncology Workgroup, and contained the data of all the patients with lymphoid leukaemia 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. They did not found any seasonal pattern and they concluded there is little evidence of any seasonality in the diagnosis of acute lymploblastic leukaemia.

To sum up, because of the importance of chronic diseases and increase the number of people who suffer from this, Abdar et al. studied liver disease through using two computational intelligence methods [Abdar et al, 17]. In the mentioned application, they considered 583 instances of liver disease dataset from the UCI repository. However, they do data pre-processing, and modify the original dataset.

It should be noted that each of the solutions described above has weaknesses and limitations, which limit the desired results, with the negative consequences in the diagnoses they produce. Therefore, it is necessary to carry out research, such as the present article, where solutions are proposed that overcome the weaknesses and limitations of current models.

3 Proposed model for medical data diagnosis

This section is devoted to the explanation of the proposed model for medical diagnosis: the Assisted Classification for Imbalance Data (ACID) model. Subsection 3.1 explains the main ideas of the model, as well as its functioning. Subsection 3.2 details the training phase of ACID, while subsection 3.3 addresses the classification phase.

ACID model is designed to deal with imbalanced data, very common in medical domains (usually, the number of sick people is much lower than the number of healthy ones). In addition, it is intended for handling mixed as well as incomplete data, which is also a common situation in medical scenarios. Since ACID is a supervised model, it requires as an input a set T of classified instances, described by a set of features $= \{A_1, \dots, A_n\}$. The value of the i-th feature of an instance $x \in T$ is denoted by x_i. If this value is unknown (missing), $x_i = '?'$. Each instance must belong to only one of a set of classification classes $K = \{K_1, \ldots, K_p\}$.

3.1 Assisted Classification for Imbalance Data (ACID)

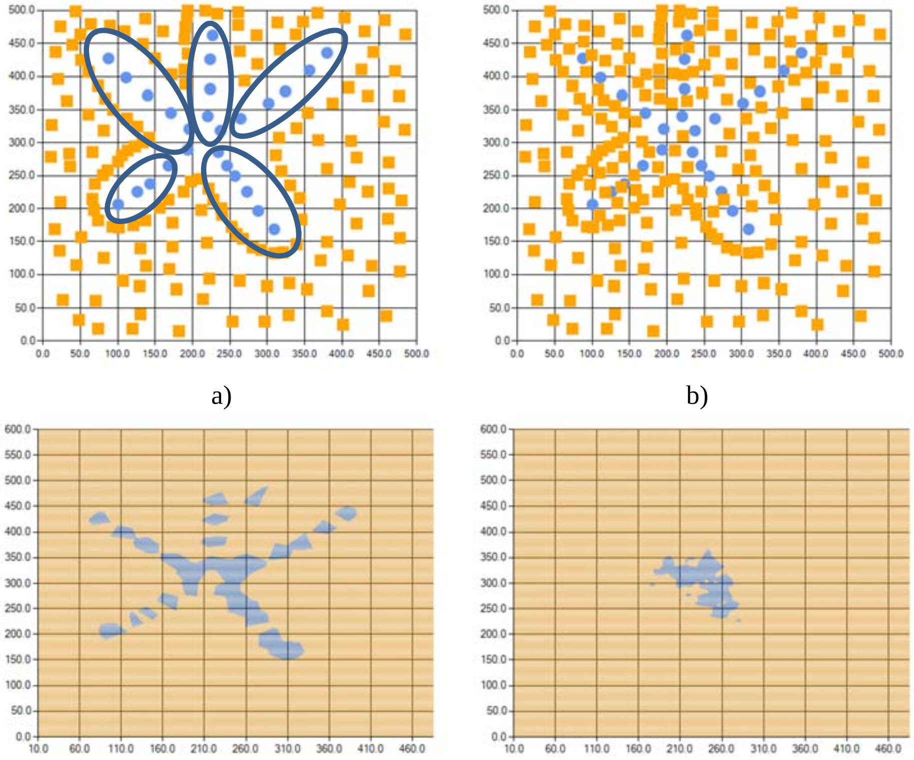
Wolpert demonstrate that no classification algorithm can outperform all others according to all performance measures, in all domains [Wolpert, 01]. However, some desired characteristics included in the proposed model makes ACID a good classification model for medical diagnosis.

One of the key aspects of medical diseases datasets is that they are often imbalanced. That is, the number of instances in the majority class is greater than the number of instances in the minority class, which is usually the class of interest. The classification of imbalanced data is challenging and several factors influence it. In the following, we discuss some of them [López et al., 13]:

1. Bias to majority class. Some classifiers bias towards majority class, due to their functioning. For instance, Neural Networks do weights adjustment considering the overall error, not the error of each of the classes. Similarly, Decision Trees usually have pruning or leave creation criteria based on the overall homogeneity of leaves, while Support Vector Machines adjust their parameters considering the overall performance.
2. Presence of small disjoints. Small disjoints are found when a class occupies different region in the feature space. Thus, the so-called subclasses or subtypes emerge. Subclasses contains instances of the same class, but very different from each other. The presence of small disjoints affects classifiers a lot.
3. Class overlapping. In such scenarios, the Bayes risk increases, and the performance of classifiers diminish.
4. The curse of dimensionality. It affects similarity-based classifiers, due to as the number of dimensions increases, the instances tend to appear to be "more similar".

In figure 1, we show an example of the clover dataset [Napierala et al., 10], exemplifying some of the above-mentioned factors influencing imbalanced data classification. Medical datasets usually have those factors. In [Napierala et al., 10], the authors show a representation of two medical datasets, in which is clear the presence of small disjoints and class overlapping. In addition, in such study the authors mention the curse of dimensionality.

Patients descriptions may be somehow dispersed, that is, very different patients have diseases. For instance, if a value of a certain exam is outer of some normal boundaries, the results indicate a disease. For instance, an adult patient having a glucose level when fasting lower than 4.0 mmol/L (72 mg/dL) or greater than 6.0 mmol/L (108 mg/dL), is considered as sick, while a patient with a glucose level between 4.0 to 6.0 mmol/L (72 to 108 mg/dL) is considered as healthy. Figure 2 depicts this example.



a)

b)

c)

d)

Figure 1: Example of factors influencing imbalanced classification in Clover dataset. a) presence of small disjoints (subclasses) within a class, b) class overlapping, c) original decision regions and d) bias towards majority class, in this example, by the 5 Nearest Neighbor classifier.

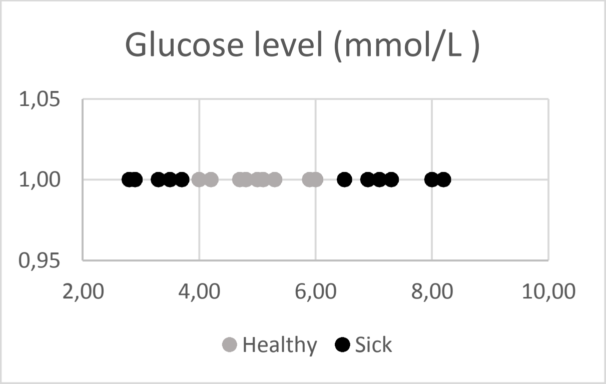


Figure 2: Example of a class (sick) occupying two regions in the feature space.

To deal with imbalanced data, affected by the previously explained factors, we design a novel classification algorithm. The ideas behind the proposed classifier are the following:

1. Diminishing the effects of the curse of dimensionality. To do so, we use a metaheuristic algorithm, to obtain attribute weights. In addition, we consider as irrelevant those attributes having lower weights. By doing so, we ensure that the classifier only considers relevant features.
2. Dealing with small disjoints. We seek for presence of small disjoints in classes. To do so, we structure each class separately, to find the different subclasses (disjoints) of the class. We consider a subclass has very similar instances, being different from the instances of other subclasses. We use clustering techniques to find those subclasses.
3. Bias to majority class. To guarantee that our classifier is unbiased towards majority class, we use a similarity-based strategy, which gives the same similarity possibilities to each class. Our strategy consists on obtaining the similarity to each subclass, and then return the most similar subclass of each class. That is, each class has only one more similar subclass. Then, we will label the instance according to the label of the most similar subclass, among all classes. It guarantees that each class has a single similarity value for comparison purposes, being unbiased to any of the classes.
4. Dealing with class overlapping. To diminish the influence of class overlapping, we aggregate the similarity of the instance to classify with respect to a subclass. That is, instead of considering the most similar instance, we use aggregation operators to compare an instance with a set of instances in a subclass. It guarantees that class overlapping and noisy or mislabeled instances influence less the classification process.

In the following, we detailed the training and classification phases of ACID. Training addresses previous ideas one and two, while classification addresses previous ideas three and four.

3.2 ACID: training phase

3.2.1 Diminishing the effects of the curse of dimensionality

The training phase of ACID starts by dealing with the curse of dimensionality. To do so, we compute attribute weights by using Differential Evolution (DE). We use Differential Evolution due to this algorithm obtained a good performance in experiments with other supervised classifiers [Ramírez et al. 15]. First, we determine the adequate codification strategy for weight computation. Features weights computation is a kind of continuous optimization problem, due to the features weights are real values.

To apply DE, we codify the feature weights as an array of real values in the [0,1] interval. It allows us to analyze the relative importance of features in the classification process. Considering this codification, zero means that the feature is completely irrelevant; while one means this feature is highly relevant for the classification task. In addition, the codification in the [0,1] interval facilitates further developments of a fuzzy version of the ACID model.

The array of values will have as length the number of features that describe the problem. Figure 3 shows an example of an individual of DE, which codifies the feature weights associated to a dataset with four attributes.

0.30	0.25	0.68	0.92

Figure 3: Codification of a Differential Evolution individual in the ACID model.

Fitness function definition constitutes a key aspect in optimization algorithms. In this research, we used averaged True Positive Rate [Fernández et al., 13] as fitness function. To obtain fitness, it is necesary to have a validation set. We divided the training set into two subsets: training and validation, by means of a Hold-Out procedure, having 70% of instances for training, and 30% of instances for fitness computation. Figure 4 illustrates fitness computation process in the ACID model.

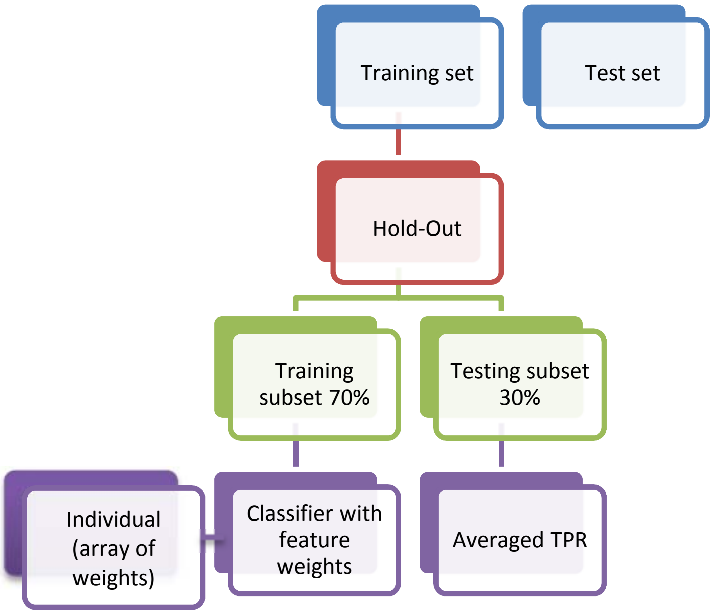


Figure 4: Fitness computation in the ACID model.

We use a population of 25 individuals and 1000 iterations. After obtained the features weights by means of Differential Evolution, ACID uses a threshold ε to delete irrelevant attributes. That is, let w_i be the weight of feature A_i, assigned by Differential Evolution. If $w_i \leq \varepsilon$, feature A_i will be considered as irrelevant and will be deleted. In the experiments, we set $\varepsilon = 0.1$.

3.2.2 Dealing with small disjoints

To deal with small disjoints, we divide the training set into classes, and then, we structure each class. Clustering algorithms aim at obtaining “the natural structure of data" [Jain and Dubes, 88]; thus, they are a useful choice to detect the presence of small disjoints in a class. However, medical data is often mixed (that is, described by both numeric and categorical attributes) and incomplete (some instances have missing attribute values). In addition, we do not know how many disjoints a class may have.

To deal with this scenario, we need a clustering algorithm able to handle mixed and incomplete data, and with no predefined number of clusters. In the Logical Combinatorial Approach to Pattern Recognition, there are several clustering procedures [Trinidad et al., 00]. One of them is the Compact Sets structuralization. Compact Sets are the connected components of a Maximum Similarity Graph (MSG).

A MSG is a directed graph that connects each instance with its most similar instances. Formally, let $G = (X, \theta)$ be a MSG for a set of instances X, where θ is the arcs set. In this graph, two instances $x, y \in X$ form an arc $(x, y) \in \theta$ if $\max_{z \in X}\{sim(x, z)\} = sim(x, y)$, where $sim(x, y)$ is a similarity function. Usually $sim(x, y) = 1 - d(x, y)$ and $d(x, y)$ is a dissimilarity function. In case of ties, the MSG establishes a connection between the instance and each of its nearest instances. Compact Sets are the connected components of such graph. Formally, a subset $N \neq \emptyset$ of X is a Compact Set if and only if [Ruiz-Shulcloper and Abidi, 02]:

$$a)\forall y \in X\left[x \in N \wedge \left(\begin{array}{c}\max\limits_{\substack{\mathrm{x} \in \mathrm{X} \\ \mathrm{x_i} \neq \mathrm{y}}}\{sim(x, y)\} = sim(x, y) \\ \vee \max\limits_{\substack{\mathrm{x} \in \mathrm{X} \\ \mathrm{x} \neq \mathrm{y}}}\{sim(y, x)\} = sim(y, x)\end{array}\right)\right] \Rightarrow y \in N$$

$$b)\ \forall x, y \in N, \exists x_1, \cdots, x_q \in N\left[\begin{array}{c} x = x_1 \wedge y = x_q \wedge \forall \mathrm{p}\ \{1, \cdots, \mathrm{q} - 1\} \\ \left[\begin{array}{c}\max\limits_{\substack{z \in X \\ z \neq x_p}}\{\mathrm{sim}(x_p, \mathrm{z})\} = \mathrm{sim}(x_p, x_{p+1}) \\ \vee \max\limits_{\substack{z \in X \\ z \neq x_p}}\{\mathrm{sim}(x_{p+1}, \mathrm{z})\} = \mathrm{sim}(x_{p+1}, x_p)\end{array}\right]\end{array}\right]$$

$c)$ Every isolated object is a Compact Set, which is degenerated.

Compact Sets have several advantages to data analysis. They do not assume any properties of data, and they do not need any parameter for their construction, except of the similarity function to compare two instances, and directly handle mixed as well as incomplete data. In addition, the instances connect only to their most similar instances in the training set, which is valuable information particularly in high Bayes risk zones.

We use Compact Sets [Trinidad et al, 00] for finding small disjoints in the classes. Compact Sets directly handles mixed and incomplete data, and need no parameter for its obtaining, that is, they do not require the number of clusters to obtain. After obtained the clusters of each class, they are stored for further use.

Summarizing, the training phase of ACID consist of two phases: diminishing the curse of dimensionality and detecting subclasses in classes. The process returns two kinds of data: feature weights, and the subclasses of each class. Figure 5 illustrates the training phase and its returning data in ACID.

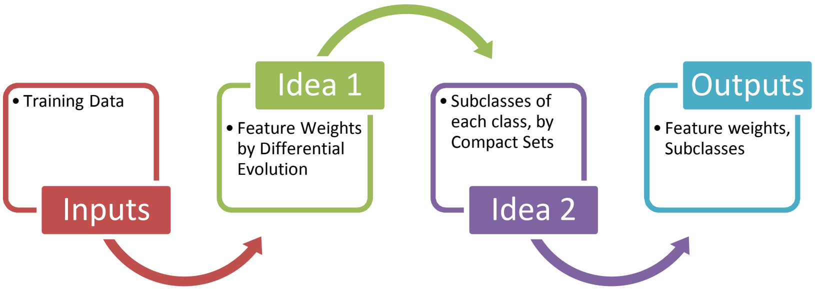

Training Data
Inputs
Idea 1
Feature Weights by Differential Evolution
Subclasses of each class, by Compact Sets
Idea 2
Outputs
Feature weights, Subclasses

Figure 5: Training phase of ACID model.

As shown, ACID computes a set of features weights $W = \{w_1, \dots, w_n\}$ associated to the feature set. For this task, ACID uses Differential Evolution [Storn and Price, 97]. In addition, in the training phase, ACID obtains a set of small disjoints of each class, which will be used later in the classification phase. Noisy and outlier instances are in the subclasses; however, they will contribute less to instance classification.

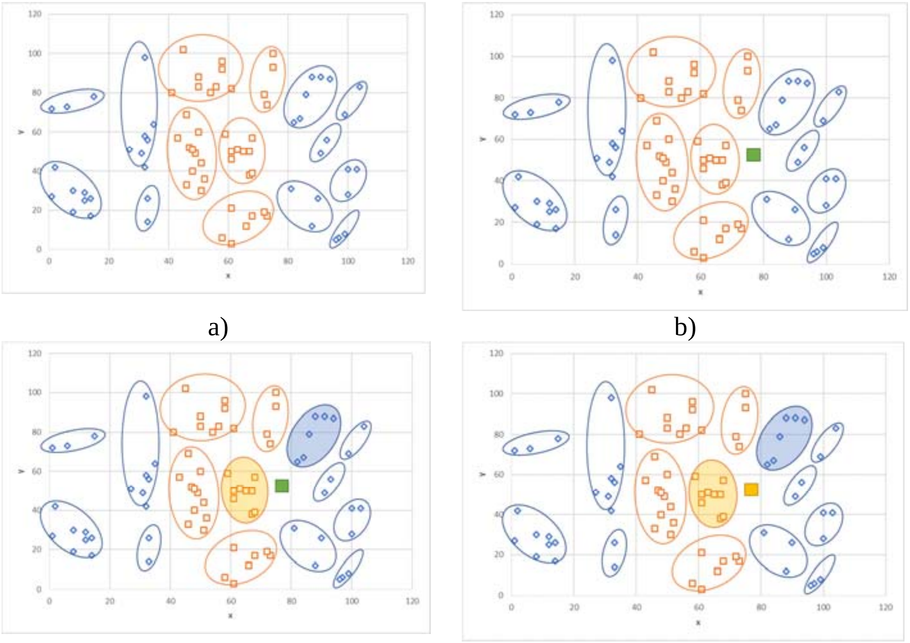
3.3 ACID: classification phase

The ACID model deals with imbalanced data and with class overlapping in the classification phase. To do so, it uses a similarity computation strategy to guarantee being unbiased to any of the classes. In addition, it uses aggregation to diminish the influence of class overlapping, as well as the one of noisy or mislabeled instances. The following subsections explain those ideas in detail.

3.3.1 Bias to majority class

To guarantee an unbiased performance, ACID model uses similarity computation. Unlike other classifiers such as Nearest Neighbor [Cover and Hart, 67], ACID computes the similarity of the instance to classify with respect to each subclass of each class. Then, it stores the most similar subclass of each of the classes. After that, ACID compares the similarity values, and labels the instance with the label of the class having the greater similarity value.

By doing so, we guarantee the selection chance to be the same for all classes. This makes ACID to be unbiased to any of the classes in the training set. Figure 6 shows an example of how ACID classifies an instance. In such example, we depicted a two-dimensional training data and the disjoints founded in the training phase. We also show the instance to classify, and the most similar disjoints for the instance. In figure 7 we summarize the classification phase of ACID.



a) b)

c) d)

Figure 6: Example of the classification phase of ACID model. a) Subclasses (small disjoints) obtained in training phase, b) unknown instance (in green) to classify, c) most similar subclasses of each class and d) assigned label (the square is now orange) according to most similar subclass.

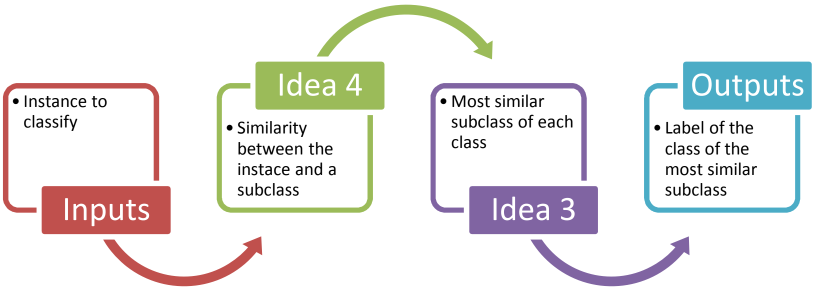

Instance to classify
Inputs
Idea 4
Similarity between the instace and a subclass
Most similar subclass of each class
Idea 3
Outputs
Label of the class of the most similar subclass

Figure 7: Classification phase of ACID model.

ACID model allows using any similarity function, or any dissimilarity function. In the later, instead of considering the most similar subclass, it considers the less dissimilar subclass to be the one containing the correct label.

3.3.2 Dealing with class overlapping

To diminish the influence of class overlapping, the ACID model considers the overall dissimilarity of the pattern to classify o, to each subclass obtained in the training phase. That is, instead of considering the most similar instance, it uses aggregation operators to compare an instance with a set of instances in a subclass. There are several measures to compute the dissimilarity between an instance and a set of instances; among the most commons are: single linkage, complete linkage, average linkage and centroid linkage [Jain, 88]. Figure 8 summarizes those measures. In the experiments, we use average linkage to determine the overall dissimilarity of an instance to a subclass.

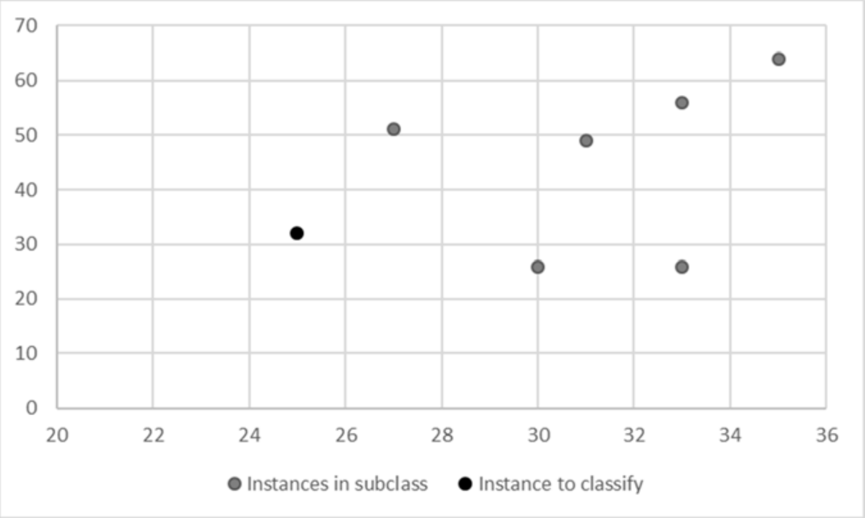
$$\begin{array}{rl}
\textit{Single-Link} & D(C_i, o) = \min\limits_{x \in C_i}\{d(x,o)\} \\
\textit{Complete-Link} & D(C_i, o) = \max_{x \in C_i}\{d(x,o)\} \\
\textit{Average-Link} & D(C_i, o) = \sum\limits_{x \in C_i} d(x,o) / |C_i| \\
\textit{Centroid-Link} & D(C_i, o) = d(\bar{x}, \bar{o})
\end{array}$$

Figure 8. Dissimilarity measures between an instance and a cluster. $\bar{x}$ *is the centroid of cluster* C_i *y* C_i *and* $d(x,y)$ *is a dissimilarity function between instances.*

Average linkage allows us to diminish the influence of noisy or outlier data in the classification process. In addition, it diminishes the influence of class overlapping. This is exemplified in Figure 9.

Summarizing, ACID is a novel classifier designed to deal with imbalanced data. It successfully handles mixed as well as incomplete datasets, and addresses some of the most relevant factors influencing imbalanced data classification. It deals with the curse of dimensionality and the presence of small disjoints; it is unbiased to any of the classes, and handles overlapped as well as noisy or mislabeled data.

In addition, ACID is an interpretable model; that is, ACID is transparent, due to we know exactly why an instance belong to a certain class. This is a clear advantage over other classifiers of the state of art.



Instance to classify: $o = (25{,}32)$

Instances in subclass:

X values	27	35	33	31	33	30	27
Y values	51	64	56	49	26	26	51
$d(x,o)$	19.10	33.53	25.30	18.03	10.00	7.81	19.10

Average-link: $D(C_i,o) = \sum_{x \in C_i} d(x,o)/|C_i| = 113.77/6 = 18.96$

Figure 9: Example of Average-link computation in ACID.

4 Experimental Results and Discussion

This section presents the experimental results obtained for medical diagnosis, using the proposed classification model, as well as other state of art models. Figure 10 illustrates the schematics of the experiment design.

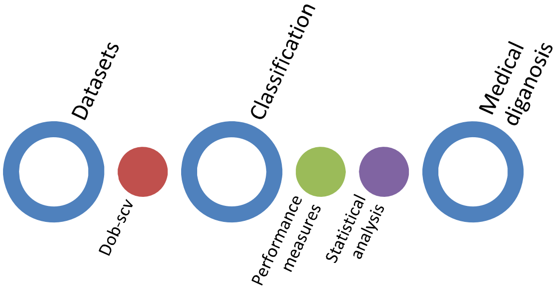


Figure 10: Schematic of the experimental design

Subsection 4.1 explains the datasets used, while subsection 4.2 describes the performance measures and statistical tests for algorithm comparisons. Subsection 4.3 analyzes the numerical results obtained by the algorithms, and subsection 4.4 discusses the statistical significance of them.

4.1 Datasets related to medical diseases

The data sets used in this research come from online information provided by both KEEL dataset repository [https://www.keel.es] and UCI dataset repository [https://archive.ics.uci.edu/ml/datasets.html]. We used the medical–related standard classification datasets. The datasets include information about different diseases, such as breast cancer, thyroid diseases, heart diseases, liver disorders, diabetes, and others. In the following, we offer a brief description of the used datasets.

Breast Cancer data set: This is one of three domains provided by the Oncology Institute. This dataset appears frequently in the machine learning literature. This data set includes 201 instances of one class and 85 instances of another class. Nine attributes describe the instances, some of which are linear and some are nominal.

Liver Disorders (BUPA) data set: This data set analyzes some liver disorders that might arise from excessive alcohol consumption (the first 5 variables), and the number of half-pint equivalents of alcoholic beverages drunk per day for each individual. The task is to select if a given individual suffers from alcoholism.

Heart Disease (Cleveland) data set: This data set is a part of the Heart Disease Data Set (the part obtained from the V.A. Medical Center, Long Beach and Cleveland Clinic Foundation), using a subset of 14 attributes. The task is to detect the presence of heart disease in the patient. It is integer valued from zero (no presence) to four.

Haberman's Survival data set: This data set contains cases from a study 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 task is to determine if the patient survived 5 years or longer (positive) or if the patient died within 5 year (negative).

Statlog (Heart) data set: This dataset is a heart disease database similar to a database already present in the KEEL repository (Heart Disease databases) but in a slightly different form. The task is to detect the absence or presence of heart disease.

Hepatitis data set: This data set contains a mixture of integer and real valued attributes, with information about patients affected by the Hepatitis disease. The task is to predict if these patients will die or survive.

Mammographic Mass data set: This data set has the task to predict the severity (benign or malignant) of a mammographic mass lesion from BI-RADS attributes and the patient's age. It contains a BI-RADS assessment, the patient's age and three BI-RADS attributes together with the ground truth (the severity field, which is the target attribute). The Institute of Radiology of the University Erlangen-Nuremberg collected the data between 2003 and 2006.

Thyroid Disease (New Thyroid) data set: This data set is one of the several databases about Thyroid available at the UCI repository [https://archive.ics.uci.edu/ml/datasets.html]. The task is to detect if a given patient is normal or suffers from hyperthyroidism or hypothyroidism.

Pima Indians Diabetes data set: From National Institute of Diabetes and Digestive and Kidney Diseases. It contains a selection of instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage. The class label represents if the person has not diabetes (tested_negative) or the person has diabetes (tested_positive).

Post-Operative data set: The classification task 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. The class label can take one of the following values: I (patient sent to Intensive Care Unit), S (patient prepared to go home), A (patient sent to general hospital floor).

South African Hearth data set: A retrospective sample of males in a heart-disease high-risk region of the Western Cape, South Africa. These data are from a larger dataset, described in Rousseauw et al, 1983, South African Medical Journal. The class label indicates if the person has a coronary heart disease: negative or positive.

SPECTF Heart data set: The dataset describes diagnosing of cardiac Single Proton Emission Computed Tomography (SPECT) images. Each of the patients belongs to one of two categories: normal or abnormal.

Thyroid Disease (thyroid0387) data set: The task is to detect is a given patient is normal or suffers from hyperthyroidism or hypothyroidism. This data set is one of the several databases about Thyroid available at the UCI repository [https://archive.ics.uci.edu/ml/datasets.html].

Breast Cancer Wisconsin diagnosis (wdbc) data set: This database contains 30 features computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image. The task is to determine if a found tumor is benign or malignant.

Breast Cancer Wisconsin original (wisconsin) data set: This dataset contains cases from a study conducted at the University of Wisconsin Hospitals, Madison, about patients who had undergone surgery for breast cancer. The task is to determine if the detected tumor is benign or malignant.

No.	Datasets	Attributes		Imbalance analysis		Missing Values	Classes
		Numeric	Categorical	Instances	IR		
1.	breast	0	9	277	2.420	Yes	2
2.	bupa	0	7	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

Table 1: Description of the datasets used.

In table 1, we offer a summary of the description of the datasets. 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 [López et al., 14], recommended for imbalanced scenarios.

4.2 Performance measures and statistical tests

Particularly, when we use imbalanced datasets for classification, the usual performance measures —such as the rate of correctly classified instances— become inadequate [Fernández et al., 13]. 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, we use the average True Positive Rate for each class as in [Fernández et al., 13].

In fact, 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 of TPR (1).

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

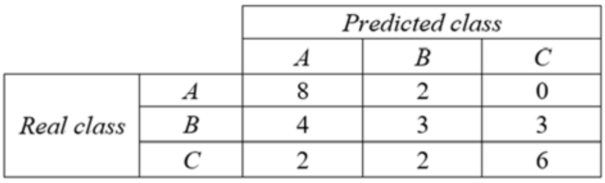
Nevertheless, in a problem with k classes, the TPR takes into consideration the total of correctly classified instances from class *j*, relative to the total of instances of the j-th class. Therefore, the TPR for class j calculate the probability of correctly classifying an instance from class j. For the computation of such measure, 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 TPR (also recall or sensitivity) of class *j*, denoted by TPR_j, is computed as follows:

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

In this research, we use the average TPR per class [Fernández et al., 13], which is defined as:

$$\overline{TPR} = \frac{1}{k}\sum_{j=1}^{k} TPR_j \tag{3}$$

In the previous equation k is the number of classes and TPR_j is the TPR 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 TPR per class allows taking into account all the classes, without bias towards any particular one. Figure 11 presents an example of how to compute average TPR, with *k* = 3 classes.



		Predicted class		
		A	B	C
Real class	A	8	2	0
	B	4	3	3
	C	2	2	6

$$TPR_A = \frac{8}{10} = 0.8, TPR_B = \frac{3}{10} = 0.3, TPR_C = \frac{6}{10} = 0.6$$

$$\overline{TPR} = \frac{0.8 + 0.3 + 0.6}{3} = \frac{1.7}{3} \approx 0.567$$

Figure 11: An example of computation of performance measures given a confusion matrix for three classes taking in consideration the average sensitivity and the minimum sensitivity.

To determine which classification algorithms got the better experimental results while diagnostic medical diseases, we use hypothesis testing. Statistical hypothesis tests evaluate whether there is a significant difference in the performance given by different classification algorithms. Regarding the works of [Demšar, 06; Garcia and Herrera, 08; Garcia et al., 10], we choose non-parametric tests for the current research. Particularly, we select the Friedman test since it is widely recommended for this kind of works~~.~~

The Friedman test [Friedman, 37; Friedman, 40] consists of ordering the samples and replacing them by their respective ranks as follows: 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 that case the test assigns an averaged rank.

If the Friedman test rejected the null hypothesis of performance equality, we need to apply a post-hoc test to determine between which algorithms there are differences [Demšar, 06; Garcia et al., 10]. Among the different post-hoc tests recommended for classification algorithms performance analysis over multiple datasets [Demšar, 06; Garcia and Herrera, 08; Garcia et al., 10] we use the Holm test [Holm, 79]. This test uses a descending (step-down) procedure to adjust the significance value α. For this, the p values are ordered ascendingly (i.e. from the most significant to the least significant). If $p_1 < \frac{\alpha}{l-1}$, the test rejects null hypothesis and the test continues the comparison with the next p value, considering whether $p_2 < \frac{\alpha}{l-2}$. This test continues this process until it cannot reject one of the hypothesis, given that $p_i \geq \frac{\alpha}{l-i}$. At this point, the test did not reject the remaining hypotheses either.

There are many specialized automated tools for the computation of the Friedman test, and the post-hoc tests. In this research, we use the KEEL software [Alcalá-Fdez et al., 09; Alcalá-Fdez et al., 11].

4.3 Numerical results obtained by the classification algorithms

Five state-of-the-art classification algorithms were selected. All of them are able to deal with mixed and incomplete data. This selection includes Nearest Neighbor (NN) [Cover and Hart, 67], Multilayer Perceptron (MLP) [Bourlard and Kamp, 88], C4.5 [Quinlan, 93], SMO [Platt, 99; Keerthi et al., 01] and Logistic [Le Cessie and van Houwelingen, 92]. For MLP, C4.5, Logistics and SMO, we used the default parameter values offered in the KEEL software package [Alcalá-Fdez et al., 09; Alcalá-Fdez et al., 11].

For the Nearest Neighbor and ACID classifiers, we use the HEOM dissimilarity [Wilson and Martinez, 97], which handles mixed and incomplete data descriptions. HEOM uses two different approach for computing dissimilarity over numeric and categorical attributes. Consider that x_a, y_a are the vales of feature a, for instances x and y, and consider that max_a and min_a are the maximum and minimum values of the feature a, the HEOM dissimilarity is:

$$
\begin{gathered}
HEOM(x,y) = \sqrt{\sum_{a=1}^{m} d_a(x_a, y_a)} \\
d_a = \begin{cases} 1 \quad \text{if } x_a \text{ or } y_a \text{ are unknown} \\ overlap(x_a, y_a) \text{ if a is categorical} \\ diff(x_a, y_a) \text{ if a is numeric} \end{cases} \\
overlap(x_a, y_a) = \begin{cases} 0 & \text{if } x_a = y_a \\ 1 & \text{othervise} \end{cases} \\
diff(x_a, y_a) = |x_a - y_a| / (max_a - min_a)
\end{gathered}
\tag{4}
$$

Table 2 shows the results obtained by the analyzed classification algorithms, for the medical diagnosis problems considered. Best results are highlighted in bold. The worst results were obtained in the *cleveland* dataset, were the best diagnosis result had an average TPR of 0.45. Another difficult to diagnosis datasets are *bupa*, *haberman*, and *saheart*, having average TPR results lower than 0.70.

The proposed algorithm obtained very good results, outperforming other classifiers in nine of the 15 datasets.

ACID obtained remarkable good results for *new-thyroid*, *wdbc* and *wisconsin* datasets, with average TPR values over 0.97. In addition, it obtained an increase in performance up to 47%, considering the second best classification algorithm. Figure 12 shows the difference in performance with respect to our proposal and the second best algorithm (values greater than zero), and with respect to the best algorithms, in the 6 datasets were our algorithm do not obtained the best results (values lower than zero).

Datasets	C4.5	kNN	Logistic	MLP	SMO	ACID
breast	0.591	0.605	0.595	0.659	0.632	**0.710**
bupa	0.614	0.652	0.659	0.535	0.500	**0.654**
cleveland	0.292	0.297	0.319	0.298	0.310	**0.405**
haberman	0.578	0.583	0.564	**0.649**	0.500	0.591
heart	0.775	0.803	0.835	0.833	0.833	**0.847**
hepatitis	0.679	0.732	0.641	0.820	0.693	**0.841**
mammographic	**0.838**	0.818	0.828	0.459	0.824	0.753
newthyroid	0.894	0.914	0.956	0.695	0.767	**0.986**
pima	0.687	0.690	**0.730**	0.708	0.714	0.700
post-operative	0.328	0.343	0.326	**0.641**	0.336	0.539
saheart	0.618	0.607	0.669	0.643	**0.688**	0.646
spectfheart	0.565	0.701	0.606	0.579	0.509	**0.762**
thyroid	**0.976**	0.593	0.724	0.447	0.518	0.756
wdbc	0.479	0.475	0.487	0.500	0.477	**0.970**
wisconsin	0.502	0.511	0.512	0.510	0.503	**0.979**
Times Best	2	0	1	2	1	**9**

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

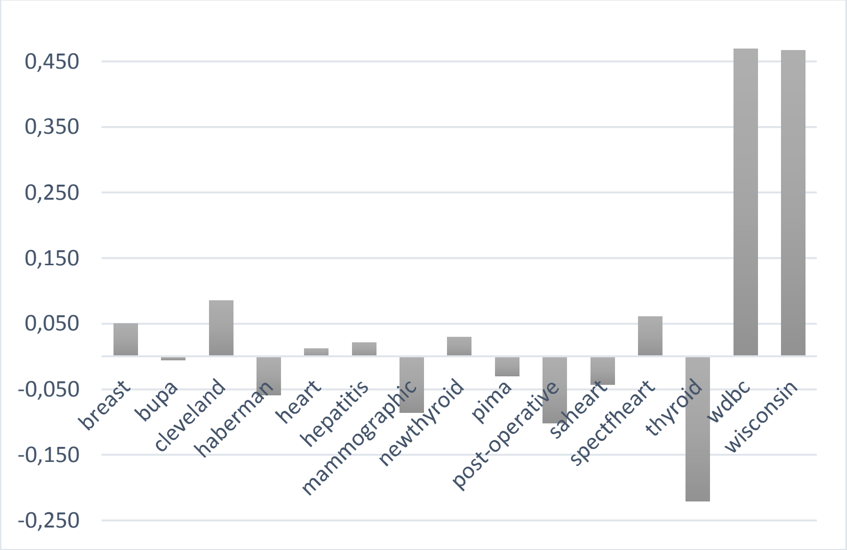


Figure 12: Difference in performance of ACID vs. other algorithms.

In order to find out if our proposal is more appropriate for the correct diagnosis of diseases, we apply the Friedman test [Friedman, 37; Friedman, 40], giving a value of $p = 0.001586$, which is largely below the established significance level of $\alpha = 0.05$ for a 95% confidence.

No.	Algorithm	Ranking
1	ACID	1.867
2	Logistic	3.000
3	MLP	3.567
4	kNN	3.933
5	SMO	4.100
6	C4.5	4.533

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

The algorithms rankings according to the Friedman test are shown in table 3, where the best classifier for this task is clearly ACID. Considering the results of the Friedman Test, we applied a post hoc test, the Holm test [Holm, 79].

The test rejects the hypothesis having an adjusted value lower or equal than 0.05. That is, it concludes that for a 95% of confidence, there are significant differences in the average TPR obtained by our proposal and every other classifier (Table 4). These results confirm that ACID classifier is adequate for medical diagnosis, with significantly better average TPR results than MLP, C4.5, 3-NN, SMO and Logistic classifiers.

i.	Algorithm	z	p	Adjusted Holm
5	C4.5	3.9036	0.000095	0.01
4	SMO	3.269265	0.001078	0.0125
3	kNN	3.02529	0.002484	0.016667
2	MLP	2.488545	0.012827	0.025
1	Logistic	1.65903	0.09711	0.05

Table 4: Post hoc comparison obtained by the Holm test.

In the analysis of diagnosis of several diseases, the ACID model obtained very good results; due to it significantly outperform other classifiers in medical scenarios. Such results support the assertion that the proposal is quite useful for medical purposes.

5 Conclusions and Future Work

In this paper, we introduced a novel classification model, designed for medical diagnosis, named ACID (Assisted Classification for Imbalance Data model), which is able to handle imbalanced data, having mixed categorical and numerical attributes, and missing values. Our model deals with the presence of small disjoints in imbalanced data.

The main idea of ACID is to structure data, and to find the most similar structure to the instance to classify. This peculiar functioning allows handling problems having the same class in different decision regions. In addition, by considering the closest structure to every class, ACID successfully deals with imbalanced data, due to each class has the same representation for classification purposes. On the other hand, by structure data, ACID reduces the influence of noisy and outlier data, facilitating the correct classification of instances.

A key aspect of ACID is that it is an interpretable model; that is, ACID is transparent, due to we know exactly why an instance belong to a certain class.

Experimental results illustrate the good performance of ACID, due to it outperforms several state of art classifiers, in nine of 15 medical datasets. According to the Friedman test, the best classifier in the experiments carried out is ACID; in addition, the post hoc Holm test concludes that there are significant differences in the average True Positive Rate obtained by our proposal and every other classifier. It confirms that ACID classifier is adequate for medical diagnosis, with better results than MLP, C4.5, 3-NN, SMO and Logistic classifiers.

As future work, we intend to apply ACID model over other repository datasets, and to carry out comparisons with respect to other supervised classifiers. In addition, considering multiclass problems as well as fuzzy versions of ACID are some lines of future work.

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

[Abdar et al., 17] Abdar, M., Zomorodi-Moghadam, M., Das, R., & Ting, I. H.: "Performance analysis of classification algorithms on early detection of liver disease. Expert Systems with Applications", 67 (2017), 239-251.

[Alcalá-Fdez et al., 09] Alcalá-Fdez, J., Sánchez, L., García, S., del Jesus, M.J., Ventura, S., Garrell, J.M., Otero, J., Romero, C., Bacardit, J., Rivas, V.M., Fernández, J.C., Herrera, J.: "KEEL: A Software Tool to Assess Evolutionary Algorithms to Data Mining Problems"; Soft Computing, 13, 3 (2009), 307-318.

[Alcalá-Fdez et al., 11] Alcalá-Fdez, J., Fernandez, A., Luengo, J., Derrac, J., García, S., Sánchez, L., Herrera, F.: "KEEL Data-Mining Software Tool: Data Set

Repository, Integration of Algorithms and Experimental Analysis Framework"; Journal of Multiple-Valued Logic and Soft Computing, 17, 2-3 (2011)., 255-287.

[Bourlard and Kamp, 88] Bourlard, H., & Kamp, Y.: "Auto-association by multilayer perceptron's and singular value decomposition"; Biological cybernetics, 59(4-5), (1998), 291-294.

[Chang et al., 06] Chang, C. C., Cheng, C. S., & Huang, Y. S.: "A Web-Based Decision Support System for Chronic Diseases"; J.UCS (Journal for Universal Computer Science), 12, 1 (2006), 115-125.

[Cover and Hart, 67] Cover, T. M., & Hart, P. E.: "Nearest neighbour pattern classification"; IEEE Transactions on Information Theory, 13(1), (1967), 21-27.

[Demšar, 06] Demšar, J.: "Statistical comparisons of classifiers over multiple data sets"; Journal of Machine Learning Research, 7 (2006), 1-30.

[Fazekas, 06] Fazekas, M.: "Analysing Data of Childhood Acute Lymphoid Leukaemia by Seasonal Time Series Methods"; J.UCS (Journal for Universal Computer Science), 12, 9 (2006), 1190-1195.

[Fernández et al., 13] 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 (2013), 97-110.

[Friedman, 37] Friedman, M.: "The use of ranks to avoid the assumption of normality implicit in the analysis of variance"; Journal of the American Statistical Association, 32 (1937), 674–701.

[Friedman, 40] Friedman, M.: "A comparison of alternative tests of significance for the problem of m rankings"; Annals of Mathematical Statistics, 11 (1940), 86–92.

[Garcia and Herrera, 08] Garcia, S., Herrera, F.: "An Extension on 'Statistical Comparisons of Classifiers over Multiple Data Sets' for all Pairwise Comparisons"; Journal of Machine Learning Research, 9 (2008), 2677-2694.

[García et al., 10] 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 (2010), 2044-2064.

[Havlik et al., 13] Havlik, J., Lhotska, L., Parak, J., Dvorak, J., Horcik, Z., & Pokorny, M.: "A Modular System for Rapid Development of Telemedical Devices"; J.UCS (Journal for Universal Computer Science), 19,9 (2013), 1242-1256.

[Holm, 79] Holm, S.: "A simple sequentially rejective multiple test procedure"; Scandinavian Journal of Statistics, 6 (1979), 65–70.

[Jain and Dubes, 88] Jain, A. K., & Dubes, R. C.: "Algorithms for clustering data"; Prentice-Hall, Inc. (1988).

[Keerthi et al., 01] Keerthi, S.S., Shevade, S.K., Bhattacharyya, C., Murthy, K.R.K.: "Improvements to Platt's SMO Algorithm for SVM Classifier Design"; Neural Computation, 13, 3 (2001), 637-649.

[Le Cessie and van Houwelingen, 92] Le Cessie, S., van Houwelingen, J.C.: "Ridge Estimators in Logistic Regression"; Applied Statistics, 41, 1 (1992), 191-201.

[Leema et al., 16] Leema, N., Nehemiah, H. K., & Kannan, A.: "Neural network classifier optimization using differential evolution with global information and back propagation algorithm for clinical datasets"; Applied Soft Computing, 49 (2016), 834-844.

[López et al., 13] 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 (2013), 113-141.

[López et al., 14] López, V., Fernández, A., Herrera, F.: ". On the importance of the validation technique for classification with imbalanced datasets: Addressing covariate shift when data is skewed"; Information Sciences, 257 (2014), 1-13.

[Napierala et al., 10] Napierala, K., Stefanowski, J., Wilk, S.: "Learning from imbalanced data in presence of noisy and borderline examples"; In Rough sets and current trends in computing. Springer Berlin/Heidelberg (2010), 158-167.

[Platt, 99] Platt, J. C.: "12 fast training of support vector machines using sequential minimal optimization"; Advances in Kernel Methods (1999), 185-208.

[Quinlan, 93] Quinlan, R.: "C4.5: Programs for Machine Learning"; Morgan Kaufmann Publishers, San Mateo, CA (1993)

[Ramírez et al. 15] Ramirez, A., Lopez, I., Villuendas, Y., Yanez, C.: "Evolutive improvement of parameters in an associative classifier"; IEEE Latin America Transactions, 13, 5 (2015), 1550-1555.

[Rijo et al., 16] Rijo, R., Silva, C., Pereira, L., Gonçalves, D., & Agostinho, M.: "Decision Support System to Diagnosis and Classification of Epilepsy in Children"; J.UCS (Journal for Universal Computer Science), 20, 6 (2014), 907-923.

[Ruiz-Shulcloper and Abidi, 02] Ruiz-Shulcloper, J. and Abidi, M.A.: "Logical combinatorial pattern recognition: A Review"; In S. G. Pandalai (Ed.), Recent Research Developments in Pattern Recognition, USA: Transword Research Networks (2002), 133-176.

[Storn and Price, 97] Storn, R. and Price, K.: "Differential evolution – a simple and efficient heuristic for global optimization over continuous spaces"; Journal of Global optimization, 11(4) (1997), 341-359.

[Trinidad et al., 00] Trinidad, J. F. M., Shulcloper, J. R., & Cortés, M. S. L.: "Structuralization of universes"; Fuzzy sets and systems, 112, 3 (2000), 485-500.

[Wilson and Martinez, 97] Wilson, D. R., & Martinez, T. R.: "Improved heterogeneous distance functions"; Journal of Artificial Intelligent Research, (1997), 1-34.

[Wolpert, 01] Wolpert, D. H.: "The supervised learning no-free-lunch theorems"; In Proceedings of the 6th Online World Conference on Soft Computing in Industrial Applications, 6, 1 (2001), 1-20.