journal homepage: www.intl.elsevierhealth.com/journals/cmpb

Computer-aided diagnosis system: A Bayesian hybrid classification method

F. Calle-Alonso, C.J. Pérez, J.P. Arias-Nicolás, J. Martín*

Department of Mathematics, Faculty of Veterinary Medicine, University of Extremadura, Avda. de la Universidad s/n, 10003 Cáceres, Spain

ARTICLE INFO

Article history:
Received 30 November 2012
Received in revised form
29 April 2013
Accepted 26 May 2013

Keywords:
Bayesian methodology
Classification
Computer-aided diagnosis
Relevance feedback

ABSTRACT

A novel method to classify multi-class biomedical objects is presented. The method is based on a hybrid approach which combines pairwise comparison, Bayesian regression and the *k*-nearest neighbor technique. It can be applied in a fully automatic way or in a relevance feedback framework. In the latter case, the information obtained from both an expert and the automatic classification is iteratively used to improve the results until a certain accuracy level is achieved, then, the learning process is finished and new classifications can be automatically performed. The method has been applied in two biomedical contexts by following the same cross-validation schemes as in the original studies. The first one refers to cancer diagnosis, leading to an accuracy of 77.35% versus 66.37%, originally obtained. The second one considers the diagnosis of pathologies of the vertebral column. The original method achieves accuracies ranging from 76.5% to 96.7%, and from 82.3% to 97.1% in two different cross-validation schemes. Even with no supervision, the proposed method reaches 96.71% and 97.32% in these two cases. By using a supervised framework the achieved accuracy is 97.74%. Furthermore, all abnormal cases were correctly classified.

1. Introduction

Pattern recognition solves the problem of assigning each input value or object to one of a given set of classes [1–3]. This research area has become very active in the last several years. Its importance has increased since the development of new Content-Based Image Retrieval (CBIR) methods and the improvement of classification techniques. The growth of information technologies has produced a huge increase of available information, especially images and videos.

Image classification methods are being used in many disciplines, but especially relevant are those applications related to health sciences. In fact, one of the most important pattern recognition application areas is diagnostic imaging in medicine [4]. Many modalities of diagnostic imaging (conventional X-ray, computed tomography, nuclear medicine, magnetic resonance imaging, ultrasound scans, etc.) are especially useful because they help to provide effective diagnoses in a noninvasive way. Besides, during the past several years there has been an important increase in the use of diagnostic medical imaging [5].

Computer-aided diagnosis (CAD) is a broad concept that integrates signal processing, artificial intelligence and statistics into computerized techniques that assist health professionals in their decision-making processes. These techniques seek to maximize the information that may be automatically extracted from medical images via objective and quantitative computations. Due to the high volume of images and the amount of information currently provided,

* *Corresponding author*. Tel.: +34 610764510.
E-mail address: fcalonso@unex.es (F. Calle-Alonso).
0169-2607/$ – see front matter © 2013 Elsevier Ireland Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.cmpb.2013.05.029

CAD systems have become powerful tools to assist physicians in achieving high efficiency [6]. They have greatly increased the knowledge of normal and diseased anatomy, and can be determinant in diagnosis and treatment planning.

CAD systems play an important role in the early detection of diseases. For example, accurate and early diagnoses of Alzheimer's disease is key for the development of effective treatments that can palliate the effects of this neurodegenerative disease. Ref. [7] presented a CAD system for the early detection of Alzheimer's disease by means of single-photon emission computed tomography. Refs. [8–10] and references therein also consider CAD systems that use different types of images and classification algorithms for early detection of Alzheimer's disease. Diagnosing osteoporosis can also be benefited from these computerized techniques. Thoracic and lumbar vertebrae are the most common sites of osteoporosis-related fractures. It is very important to detect vertebral fractures as early as possible because timely pharmacologic intervention can reduce the risk of subsequent additional fractures. Ref. [11] developed a CAD system to detect vertebral fractures by using lateral chest radiographs. Ref. [12] retrospectively evaluated the usefulness of CAD systems to radiologist performance in the detection of vertebral fractures and lung nodules on chest radiographs. There are many other diseases that CAD systems may help to detect. In fact, they can help to diagnose any disease that can be detected from images or other biological signals by automatically extracting features [13–15].

A broad vision of CAD is not limited to image analysis. For example, certain specific characteristics obtained from voice recordings can be useful in diagnosing speech disorders. Numerous techniques for automatic evaluation of speech disorders have been proposed in the last several years [16]. Voice recordings have also been used to diagnose Parkinson's disease. Ref. [17] developed new speech signal processing algorithms to classify Parkinson's disease patients. The Parkinson's Voice Initiative[1] has opened the possibility to advance in the diagnosis of the disease by using a large database obtained by automatized telephone calls.

Both the feature extraction processes (pre-processing) and the classification techniques are currently challenges for the development of efficient CAD systems [15]. The pre-processing step provides the necessary ingredients to apply the classification algorithms. Sometimes dimensional reductions are performed [18,19]. There is a wide variety of available classification methods that have been used for CAD systems, for example, linear discriminant [20], logistic regression [21], Support Vector Machine (SVM) [22], segmentation techniques [23], k-nearest neighbor [24], and artificial neural networks [25], among others. Sometimes conceptually simple algorithms are considered, whereas other times more sophisticated methods are developed. However, there is no global best classifier, but some work better than others in specific contexts. The important issue is that the classifier can be properly integrated in a CAD system and that high levels of accuracy are obtained.

In spite of the advances in this field, there are still few powerful methods that consider learning processes and even fewer that apply the Bayesian methodology in a relevance feedback framework. The power of this methodology remains to be exploited because learning processes can be determinant for the classification results. Most of the developed Bayesian classification methods have been focused on binary classification, although multi-class problems have also been tackled. Ref. [26] classified patients in two classes (with or without Alzheimer's disease), using a Bayesian classifier without learning. The number of available objects was much lower than the dimension of the feature space, so first of all they applied a principal component analysis. Ref. [27] classified renal cell carcinoma (four classes) with a Bayesian approach without learning. Ref. [28] use several classification and regression methods, including Bayesian networks, and meta-learning algorithms such as bagging [29] and AdaBoostM1 [30]. These methods have been applied for pancreatic cancer detection in a multi-class setting, showing that Bayesian techniques provided the best overall performance. Ref. [31] test the Bayesian classifier versus classical logistic regression, finding out that the former performs better than the latter for automatic classification of polycystic ovary syndrome (two classes). Ref. [32] classified patients with and without depressive disorder by using neuroimaging scans with two machine learning techniques: relevance vector machine (Bayesian) and SVM (classical).

In this work, a novel Bayesian hybrid classification method that can be used with and without relevance feedback is presented. This method is specially indicated to classify biomedical objects containing a high number of characteristics and where there are not many objects in each class. Cancer classification problems are one of the interesting applications. By combining the power of three different techniques (pairwise comparison, Bayesian regression, and k-nearest neighbors (KNN)), this method provides a high accuracy in classification. The method can run in a fully automatic way or it can incorporate relevance feedback. In the latter case, the information obtained from both an expert and the automatic classification is used to improve the results through a relevance feedback framework. Although the applications presented in this paper focus on two concrete medical diagnostic problems, the approach is useful for many other purposes. The only limitation is that the features of the biomedical objects must have enough information to allow a discrimination.

The outline of the rest of the paper is as follows. Section 2 presents the approach by including a general description, the information about the features, the algorithm involved and the relevance feedback. In order to illustrate the applicability of the method, two different problems in the area of medical diagnosis are presented in Section 3. Finally, some conclusions are summarized in Section 4.

2. The approach

A novel hybrid classification approach is proposed. The interest is focused on classifying objects in several classes according to a similarity criterion.

The method can be summarized as follows. Firstly, a set of correctly classified objects are considered as a training dataset. The objects have features, represented by numerical vectors, that have been previously extracted. Then, a pairwise

[1] http://www.parkinsonsvoice.org/.

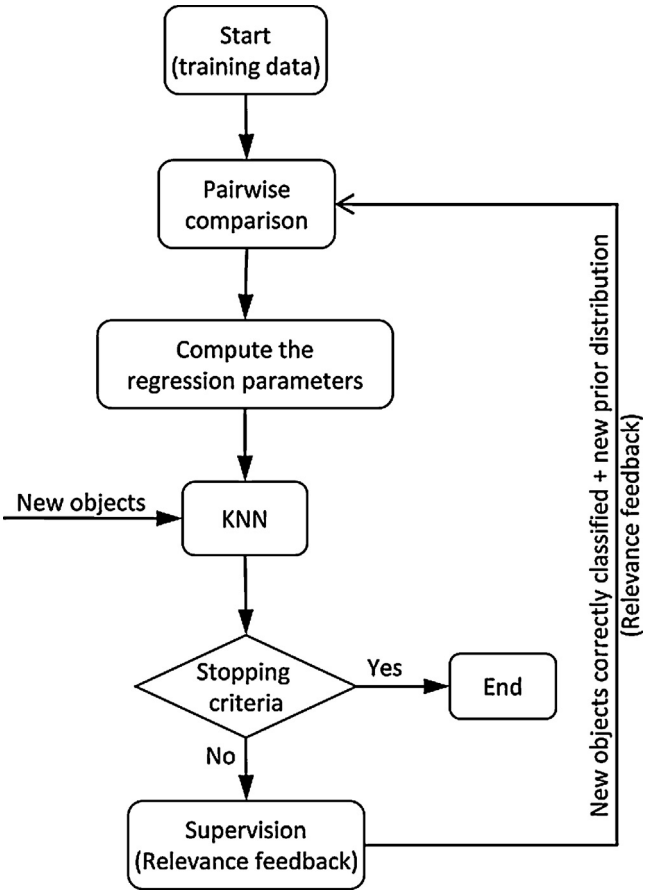

Fig. 1 – Flow diagram of the complete process.

comparison is executed and the matrix of differences is constructed. Next, the Bayesian binary regression is applied to obtain information for the classification rule in terms of regression parameters. With these current parameters and the KNN method, new objects are classified. If no relevance feedback is considered, the application of the method has finished. Otherwise, an expert supervises the classification results relocating the wrong assignments. The information provided by the expert will be used in the next iteration. Also, the posterior distribution for the parameters in the current iteration is used as the new prior distribution for the parameters in the next iteration. A continuous adaptive learning process can be applied until the method is able to classify new objects in an automatic way with a prefixed success rate.

A flow diagram of the complete process is represented in Fig. 1. The different blocks will be explained in detail in the following subsections. The whole process is implemented in R programming language.

2.1. *Feature extraction*

Feature extraction is an essential pre-processing step to perform classification based on pattern recognition. Each classifiable object is represented by a numerical feature vector extracted from it. The features can be continuous, discrete or binary data.

Medical diagnosis is often based on images (radiographies, ecographies, etc.). The information extracted from these images can be useful to classify patients into groups that fulfill any determined condition. For example, a tool integrated in *Qatris iManager* [33] can be used to extract color, texture and shape features from images. Sometimes the features are directly obtained from measurements (temperature, kind of tissue, age of the patient, etc.) and not from images. The objects to classify must be defined by any assortment of features, what is important is that they are able to describe the characteristics of the objects properly.

Assume that r objects with true classification (training dataset) are considered. Each object is represented by an $M-$dimensional feature vector and the r feature vectors are denoted by $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_r$. They will be used in the following subsection to build the matrix of differences.

2.2. *Pairwise comparison*

The pairwise comparison method was first proposed in 1927 [34]. It is based on comparing all the objects pairwise to determine if they are similar or not by using a fixed difference function. Ref. [35] proposed a pairwise comparison model based on logistic regression in a group decision making context. This method was adapted to build a CBIR system in the context of bronchoscopy analysis [36]. Now, the pairwise comparison model proposed in [35] is being used as part of the proposed classification approach.

The feature vectors representing the r correctly classified objects (training dataset) are compared two by two by using some specific distance measures (difference functions). Let d_m be a difference function between objects for the mth feature, with $m = 1, \ldots, M$. For each pair of objects defined by their feature vectors $(\mathbf{a}_i, \mathbf{a}_j)$, the following vector is defined:

$$\mathbf{x}_{\mathbf{a}_i\mathbf{a}_j} = (d_1(\mathbf{a}_i, \mathbf{a}_j), d_2(\mathbf{a}_i, \mathbf{a}_j), \ldots, d_M(\mathbf{a}_i, \mathbf{a}_j)).$$

Besides, a binary variable is also defined:

$$y_{\mathbf{a}_i\mathbf{a}_j} = \begin{cases} 0 & \text{if the objects belong to the same class,} \\ 1 & \text{if the objects belong to different classes.} \end{cases}$$

The vectors $\mathbf{x}_{\mathbf{a}_i\mathbf{a}_j}$ and the variables $y_{\mathbf{a}_i\mathbf{a}_j}$ are going to play the role of covariates and response variable, respectively, in the Bayesian binary regression model that will be described in the next subsection. The matrix of differences $\boldsymbol{\Lambda}$ is defined as:

$$\boldsymbol{\Lambda} = \begin{pmatrix} y_{\mathbf{a}_1\mathbf{a}_2} & y_{\mathbf{a}_1\mathbf{a}_3} & \cdots & y_{\mathbf{a}_{r-1}\mathbf{a}_r} \\ 1 & 1 & \ldots & 1 \\ d_1(\mathbf{a}_1, \mathbf{a}_2) & d_1(\mathbf{a}_1, \mathbf{a}_3) & \ldots & d_1(\mathbf{a}_{r-1}, \mathbf{a}_r) \\ d_2(\mathbf{a}_1, \mathbf{a}_2) & d_2(\mathbf{a}_1, \mathbf{a}_3) & \ldots & d_2(\mathbf{a}_{r-1}, \mathbf{a}_r) \\ \vdots & \vdots & \ldots & \vdots \\ d_M(\mathbf{a}_1, \mathbf{a}_2) & d_M(\mathbf{a}_1, \mathbf{a}_3) & \ldots & d_M(\mathbf{a}_{r-1}, \mathbf{a}_r) \end{pmatrix}.$$

This matrix contains the values of the response variable and the differences between features for all the possible combinations of objects without repetition. Each row is a vector of dimension $r(r-1)/2$, i.e. $(d_m(\mathbf{a}_1, \mathbf{a}_2), d_m(\mathbf{a}_1, \mathbf{a}_3), \ldots, d_m(\mathbf{a}_1, \mathbf{a}_r), d_m(\mathbf{a}_2, \mathbf{a}_3), d_m(\mathbf{a}_2, \mathbf{a}_4), \ldots, d_m(\mathbf{a}_2, \mathbf{a}_r), \ldots, d_m(\mathbf{a}_{r-1}, \mathbf{a}_r))$. Besides, a

row of 1's has been added to consider a regression model with intercept. Later, a Bayesian binary regression will be applied with the information of $\boldsymbol{\Lambda}$ [37].

2.3. Bayesian regression

Predictive models are used in a variety of medical domains for diagnostic tasks. In most cases, they are built from experience, which constitutes a valuable information acquired from the past. The Bayesian methodology has the benefit to include all the available information through the so called prior distribution of the model parameters. This methodology is considered in the proposed approach through a Bayesian binary regression model.

A binary regression model is considered, so the independent variable follows a Bernoulli distribution $y_{a_i a_j} \sim Bernoulli(\pi_{a_i a_j})$, where π is defined as:

$$\pi_{a_i a_j} = F(\boldsymbol{\beta}^T \mathbf{x}_{a_i a_j}), \quad (1)$$

with F being a cumulative distribution function (cdf), and $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_M)^T$ the regression parameter vector. Note that Bayesian regression models consider the regression parameters as random variables. The uncertainty about $\boldsymbol{\beta}$ is characterized through a probability distribution.

The probit regression model is obtained when F is the standard normal cdf and the logit regression one when F is the logistic cdf. Several regression models (probit, logit, cloglog, scobit, skew probit, etc.) can be considered. In order to select the specific cdf, a model choice approach can be used (e.g., the Deviance Information Criterion (DIC), [38]). However, in general, very small differences are obtained in this context when estimating the necessary probabilities to apply the KNN method, so the simplest model should be used. Note that in classical statistics, the most used binary regression model is the logistic one because the odds ratios are easily interpretable. However, in this Bayesian context the odds ratios are not relevant, and the computational efficiency plays the major role. The probit model [37] is the most efficient one for computational purposes. Therefore, it is proposed to use it, avoiding to perform a model choice.

In this context, the lower the value of the difference function between objects $\mathbf{a}_i$ and $\mathbf{a}_j$, the lower the probability $\pi_{a_i a_j}$. This means that when $\pi_{a_i a_j}$ approaches zero, the objects $\mathbf{a}_i$ and $\mathbf{a}_j$ are more similar and it is more probable that they belong to the same class.

Firstly, the regression method is applied to the objects from the training dataset with a weakly informative prior distribution for the regression parameters [39]. Specifically, a normal prior distribution with mean equal to zero and high variance (to let the parameters vary in a large range) is considered. The reason is that the first time the regression is applied, there is no information about the parameter distributions. Then, the posterior distribution of the regression parameters must be estimated through numerical methods. Generally, Markov Chain Monte Carlo methods (MCMC) are used for this kind of models. Specifically, a Gibbs sampling algorithm has been implemented in R, following the proposal in [37]. This step provides the parameter estimations. After these estimations have been obtained, new objects are compared with the objects in the training dataset. Consider a new object $\mathbf{a}_{new}$. The pairwise comparisons between $\mathbf{a}_{new}$ and any other $\mathbf{a}_i$ are performed. Then, $\pi_{a_{new} a_i} = \Phi(\boldsymbol{\beta}^T x_{a_{new} a_i})$ are obtained and these probabilities are used in the KNN method that will be described in the next subsection.

2.4. KNN algorithm

The Bayesian binary regression only provides information on whether the objects are similar or not (the lower the probabilities, the more similar the objects). A class assignment is performed by using the KNN multi-class algorithm. An unclassified new element in the collection is assigned to the class represented by a majority of its k-nearest neighbors in the training dataset.

This algorithm is a non-parametric procedure introduced in [40], frequently used in the pattern recognition literature. Ref. [41] provided a statistical justification and improvement of this procedure.

The classification of new objects (based on the KNN approach) is applied as follows:

1 Measure the differences between the new objects and each one of the training dataset.
2 Obtain $\pi_{a_{new} a_i} = \Phi(\boldsymbol{\beta}^T x_{a_{new} a_i})$.
3 Select the k-nearest neighbors for every new object, i.e., the k lowest probabilities obtained between the new object and the ones in the training dataset $\pi_{a_{new} a_i}$.
4 Classify each new object in the class which appears most frequently in the subset of neighbors.

2.5. Relevance feedback

Learning procedures have been studied from various viewpoints for more than fifty years since Rosenblatt proposed the Perceptron [42]. Although various types of algorithms for learning systems with a supervisor have been proposed, they can be resumed into one single class called error correction procedures. Here the interest is focused on this kind of learning processes, with an expert supervising and correcting the possible classification errors at the initial stages (learning phase).

The learning process begins when some new objects are automatically classified by using the current regression parameters. They can be assigned to the correct class or not, because the system is not free of errors. In order to improve the classification and provide high quality results, an expert supervises the new objects that have just been automatically classified. At the end of each iteration, the model receives the belonging classes marked by the expert and the features of the corrected objects jointly with their classes will be used in the next iteration. In order to do this, first all the possible pairs between the initial objects and the new ones are considered, so new columns are added to the matrix $\boldsymbol{\Lambda}$ (as many columns as comparisons between the new elements and the old ones). Every new column includes $y = 0$, if the two compared objects are in the same class, or $y = 1$, if they belong to a different one. The differences are also incorporated. This information will be very important in the next steps because it will adapt the model parameters to fit the real classification of the elements.

The first time the parameter estimation is performed (based on the training dataset), a weakly informative prior distribution is used. Then, the posterior distribution of the regression parameters is obtained for the current iteration. This posterior distribution will be used as the new prior distribution for the regression parameters in the next iteration. The prior distribution will be updated at every iteration, so that the model "learns" from the past experiments. The interactive learning process allows to update the model parameters $\boldsymbol{\beta}$ in a continuous way.

Both the information provided by the new objects supervised by the expert and the new information from the posterior distribution will continue improving the results until the process has learnt enough to be applied in a completely automatic way. Then, the learning phase is finished and the classifier is ready to be used without human interaction.

3. Applications

In this section, two classification experiments with previously published biomedical datasets (UCI Machine Learning Repository, [43]) are performed. The first one has a comparative purpose and no relevance feedback is applied. The second one shows how the proposed method performs when information can be used in a relevance feedback framework.

3.1. *Breast cancer diagnosis*

A real experiment in the context of cancer diagnosis is performed. The classification problem is based on [44]. In this publication, the authors classify 106 patients into 6 classes by using the results of an Electrical Impedance Spectroscopy (EIS) test. This type of test to diagnose breast cancer has many advantages, such as being a minimally invasive technique, very easy to use and also inexpensive. EIS applies clinically relevant frequencies to the tissue to obtain some cellular properties, such as amount of intracellular and extracellular water, packing, density and shape [45]. It has been used to detect some different diseases, for example, atherosclerotic lesions [46], malignant melanoma [47] or prostate cancer [48]. Concerning breast cancer, [49] and, recently, [50] found significant differences in breast tissue by using this technique, so breast cancer could be fairly detected by EIS.

In this experiment, breast tissue was sampled from 106 patients undergoing breast surgery. The nine features used were extracted from the Argand plot with a truncated spectra consisting of the upper seven points [44]. Four of the features were previously defined and statistically studied in [51]. These features are: impedivity at zero frequency, phase angle at 500 kHz, high frequency slope of phase angle (at 250, 500 and 1000 kHz), and impedance distance between spectral ends. The other five features, newly presented in [44], are: area under spectrum, area normalized by impedance distance between spectral ends, maximum of spectrum, distance between impedivity at zero frequency and real part of the maximum frequency point, and length of spectral curve.

There are six tissue classes: (1) connective tissue, (2) adipose tissue, (3) glandular tissue, (4) carcinoma, (5) fibroadenoma, and (6) mastopathy. The first three are normal and the others are pathological. The goal is to discriminate all of them. However, special attention must be paid to carcinoma. Ref. [44] used a method based on linear discriminant analysis. With this classic method the dimension is very important because overfitting could happen and the final classification would be affected. As a consequence, a previous reduction of the number of features is required in some cases. This inconvenient does not occur with the method proposed here, in which all the features can be included and contribute to provide information.

They found that it was very difficult to solve the problem with six classes in one step. Therefore, they proposed to use a two-stage hierarchical approach [52]. In the first stage, two groups of classes were considered: fatty tissues (adipose and connective) and the other four classes taken together. This first stage is simply a step in the hierarchical approach. It does not constitute an interesting classification problem in itself because it does not discriminate pathological classes. In the second stage, this four-class group was split into the following three categories: carcinoma, fibroadenoma + mastopathy, and glandular tissue. The authors found it very hard to discriminate between fibroadenoma and mastopathy by using the available features. They argued that the most dangerous case is the incorrect discrimination of carcinoma, so they also considered a two-class problem (carcinoma against fibroadenoma +mastopathy + glandular tissue).

In order to make a fair comparison between methods, the same conditions as in [44] are considered. A stratified cross-validation procedure has been performed for the proposed method, logistic regression, and Naive Bayes. The training dataset consists of one half of the cases, randomly selected among groups by a stratified sampling. Later on, the proposed classifier is tested by using the other half (test set). Each experiment is repeated ten times and the results are averaged. The global accuracies and the accuracies for each class are provided jointly with the 95% confidence intervals obtained from the ten iterations ($\bar{x} \pm 1.96\, s/\sqrt{10}$). The methods have been applied by using all the features ($M = 9$). Since no initial information is available, the Bayesian probit binary regression model used a weakly informative Gaussian distribution. In the KNN method, $k = 10$ is used.

Firstly, the two-class problem is considered, i.e., discrimination of carcinoma from fibroadenoma + mastopathy + glandular. Table 1 shows the global accuracies and the accuracies obtained for each class jointly with the 95% confidence intervals. Ref. [44] reported neither confidence intervals nor the percentages to calculate them. Nevertheless, note that the accuracies obtained in [44] are much smaller than the lower limits of the 95% confidence intervals obtained by using the proposed method. Even more, for logistic regression and Naive Bayes methods the upper limits of the confidence intervals for the global accuracy are smaller than the lower limit of the confidence interval provided by the proposed method. Although the results are conclusive, the percentages obtained in the iterations are used to provide quantitative information by using significance tests. Due to the small number of iterations, i.e., 10, nonparametric statistical tests are used to compare the results provided by the proposed method, logistic regression and Naive Bayes. Specifically, Kruskal–Wallis' test and the

Table 1 – Accuracies with 95% confidence intervals.

	Proposal	[44]	Logistic Reg.	Naive B.
Two classes				
Carcinoma	93.64 ± 3.56	86.36	79.22 ± 9.28	90.98 ± 4.08
Fib + Mas + Gla	99.20 ± 1.05	94.54	89.26 ± 4.69	91.68 ± 3.35
Global accuracy	97.50 ± 1.56	92.21	86.00 ± 2.82	91.43 ± 2.64
Three classes				
Carcinoma	93.64 ± 4.93	81.82	83.22 ± 6.26	87.38 ± 5.20
Fib + Mas	97.06 ± 2.31	75.76	63.66 ± 7.24	62.07 ± 7.27
Glandular	73.33 ± 12.14	63.64	40.17 ± 8.20	67.60 ± 16.81
Global accuracy	89.72 ± 2.41	N/A	65.14 ± 4.48	70.57 ± 3.92
Six classes				
Carcinoma	96.36 ± 2.97	81.82	71.78 ± 10.09	77.97 ± 7.58
Fibroadenoma	67.14 ± 14.34	66.67	44.45 ± 12.80	55.35 ± 13.78
Mastopathy	63.33 ± 16.38	16.67	41.19 ± 9.92	30.99 ± 9.96
Glandular	37.33 ± 9.83	54.54	59.83 ± 7.50	43.79 ± 15.62
Connective	91.43 ± 4.57	85.71	78.49 ± 13.87	84.46 ± 11.50
Adipose	91.82 ± 1.78	90.91	80.20 ± 10.94	81.73 ± 8.68
Global accuracy	77.35 ± 4.99	66.37	62.88 ± 5.15	62.83 ± 5.46

subsequent corrected Mann–Whitney U pairwise comparison tests (Bonferroni-type corrected version of Mann–Whitney U test) are considered. All p-values were smaller than 0.001. The percentages obtained by using the proposed method were significantly larger than the ones obtained by the other two methods. Naive Bayes provided larger percentages than the logistic regression.

The results obtained for the three classes discrimination (carcinoma, fibroadenoma + mastopathy, and glandular) are in the same direction as in the previous case (see Table 1). Although the overall efficiency is not shown in [44], it is obviously lower than the one provided by the proposed method. The proposed method also gives better results for each class. Kruskal–Wallis' test gives a p-value lower than 0.001, and the percentages obtained by using the proposed method were significantly larger ($p < 0.001$) than the ones obtained with the logistic regression and the Naive Bayes methods. The percentages provided by logistic regression and Naive Bayes have no statistically significant differences, $p = 0.14$.

Finally, the most difficult classification problem is considered, i.e., the one with six classes. The results in Table 1 show that the proposed method outperforms the other three methods concerning overall efficiency. Besides, the method provides a very good accuracy for carcinoma class, which is one of the main objectives. As in the previous cases, Kruskal–Wallis' test provided a significant result ($p < 0.01$), whereas the percentages obtained by using the proposed method were significantly larger ($p < 0.05$) than the ones obtained with the logistic regression and the Naive Bayes methods. The percentages provided by logistic regression and Naive Bayes have no statistically significant differences, $p = 0.6$.

Besides the accuracy, another advantage of this new approach is that it can use all the discriminating features ($M = 9$). In [44], the authors had to select a lower number of features, because linear discriminant classifiers have a restriction concerning the dimensionality ratio (expressed by the number of cases corresponding to the smallest populated class divided by the number of features). This ratio should be three or more in order to guarantee acceptable classifier reproducibility. However, this reduces the available information. The proposed method does not have this limitation.

3.2. *Column pathology diagnosis*

This situation corresponds to another type of classification problem where an expert can temporarily review and update the classification [53]. In this case, the proposed method with relevance feedback is applied to classify patients with or without vertebral column diseases. Ref. [54] used this dataset to test the accuracy of their classification method.

Two pathologies of the vertebral column are analyzed by using some specific features directly obtained from sagittal panoramic radiographies of the spine. The objects of the dataset are classified as normal or abnormal, depending on if they have a pathology or not. A number of 310 patients, including 100 normal and 210 abnormal, is considered. Within the abnormal situation there are 150 patients with spondylolisthesis and 60 with disc hernia. The six features computed are biomechanical attributes of the spino-pelvic system: angle of pelvic incidence, angle of pelvic tilt, lordosis angle, sacral slope, pelvic radius and grade of slipping. The effective use of these features for prediction of vertebral column pathologies was proposed in [55].

Some medical datasets are hard to automatically classify. A possibility is to reject the objects with little certainty on their classification and let the doctor classify them correctly. The amount of doubtful selected elements depends on the cost of erring the diagnosis. Ref. [54] apply three reject option methods. The first one uses two independent classifiers. Each classifier is specialized in one class and the classification is provided only when the probability of the object is high. The second method is a single standard binary one which depends on high posterior probabilities to classify. The last one is an extension of [56] that considers a single classifier with embedded reject option. They define C_{low} as the cost of classifying a patient as rejected, and C_{high} as the cost of erring with the automatic classification. Then, the normalized rejection cost is $w_r = C_{low}/C_{high}$ and the empirical risk to be minimized is $w_r R + E$, where R and E are the rejection and misclassification rates, respectively. They consider three values of w_r: 0.04, 0.24,

Table 2 – Accuracies achieved by the proposed method, logistic regression, Naive Bayes, and other methods evaluated in [54].

Train	Proposal	Reject \ w_r	0.04	0.24	0.48	Others	
		RejoSVM	96.5	87.9	83.5	Linear SVM	85.0
40%	96.71 ± 0.57	One classifier	96.7	87.7	82.1	KMOD SVM	83.9
		Two classifier	96.2	86.0	76.5	Logistic Reg.	84.7
						Naive Bayes	77.8
		RejoSVM	96.9	89.1	85.2	Linear SVM	84.3
80%	97.32 ± 0.44	One classifier	97.1	88.8	84.4	KMOD SVM	85.9
		Two classifier	96.6	86.3	82.3	Logistic Reg.	85.0
						Naive Bayes	77.8

and 0.48. As the rejection cost decreases, more patients are rejected from the automatic classification, so the physician has to perform more diagnostics. Obviously, the accuracy rate for the classification system increases. They compare these rejection methods with two different SVM algorithms, one with a linear kernel and another one with the KMOD kernel [57].

An alternative to the reject option methods is the proposed Bayesian learning method. It is used to train the system with an expert's opinion and also with the information of the past classifications (included in the prior distribution). Here, the direct classification method is initially applied, and later, the method with relevance feedback is used.

Following the same two cross-validation schemes defined in [56], a direct classification without relevance feedback is firstly performed. In the first one, 40% of the dataset is used as training data, and the rest is used to test the performance. In the second case, 80% is used for training and the rest to test. In order to consider the random variability due to the samples, stratified random sampling is performed. Each experiment is repeated 100 times to obtain stable results, and then the average accuracy is calculated. Table 2 shows the results. The accuracies of the proposed approach without relevance feedback are presented in the second column with the 95% confidence intervals ($\overline{x} \pm 1.96\, s/\sqrt{100}$). Columns 4–6 show the results for the reject option methods with different configurations [54]. In the last column, the results with SVMs [54], logistic regression and Naive Bayes methods are presented.

When the classification without relevance feedback is performed by sampling 40% of the dataset for training and 60% for testing, the proposed method reaches 96.71% average accuracy and the 95% confidence interval given by (96.14, 97.28). By using 80% of the dataset for training and 20% for testing, the average accuracy improves even more up to 97.32%, with 95% confidence interval given by (96.88, 97.76). The proposed method performs considerably better than the other methods. Without considering reject option methods, SVM methods reach 85.9% maximum accuracy against 97.32% obtained with the proposed method. Considering the reject option methods, only the one-classifier approach achieves the same accuracy when $w_r = 0.04$. However, this means that rejection can be done with a very low cost, so the physician has to individually make many diagnoses. The rejected objects are excluded from the classification and thus the success rate increases. As the rejection cost w_r decreases, the accuracy rises because the remaining elements are the ones which have high probability to belong to a class.

Although the results are clearly better for the proposed approach, more quantitative information can be obtained by using significance tests. Since the original iterations, from which the average accuracies [54] were obtained, are not available, a comparison among the proposed approach, logistic regression, and Naive Bayes is performed. In the first experiment, ANOVA conditions were met and statistically significant differences were detected among the three accuracies provided by the different methods ($p < 0.001$). All Tukey's pairwise comparisons were statistically significant ($p < 0.001$). In the second experiment, non-parametric tests as the ones in Section 3.1 were applied, because ANOVA applicability conditions were not met. Specifically, the homoscedasticity condition could not be assumed. There were statistically significant differences among the variances of the percentages across the three methods ($p < 0.001$). Kruskal–Wallis and corrected Mann–Whitney's tests provided p-values smaller than 0.001. The proposed method achieved the largest accuracy (96.71 ± 0.57, and 97.32 ± 0.44), followed by the logistic regression method (84.76 ± 0.37 and 85.03 ± 0.81), and, finally, the Naive Bayes method provided the smallest accuracy (77.81 ± 0.35, and 77.79 ± 0.86).

A remarkable result is obtained for both cross-validation schemes: there are no false negatives in any of the 100 random iterations. This means that none of the patients with a pathology has been misclassified. This is extremely important for medical diagnosis [58]. A false positive (a normal patient classified as abnormal) can be assumed and then diagnosed as normal with some additional tests, but a false negative has a higher health cost including in some cases the patient's life. This does not happen with the other classifiers. The specificities provided by the proposed approach are 0.8918 and 0.9180, respectively. For the first experiment, the sensitivity and specificity are 0.7390 and 0.8602 for the Naive Bayes method, and 0.8813 and 0.7770 for the logistic regression. For the second experiment, the sensitivity and specificity are 0.7379 and 0.8620 for the Naive Bayes method, and 0.8774 and 0.7935 for the logistic regression.

These results can be improved by using relevance feedback. Again, in this case, stratified random sampling is considered. Each experiment is repeated 100 times to obtain stable results, and then the average accuracies and the 95% confidence intervals are calculated. Besides the proposed method, logistic regression and Naive Bayes methods have been applied.

The first step considers some well classified data (40%) to start training. Then some more patients (20%) are classified by

Table 3 – Accuracy achieved with relevance feedback.

Step	Train	Test	Method	Accuracy
1	40%	20%	Proposal	96.82 ± 0.63
			Naive Bayes	76.81 ± 0.91
			Log. Reg.	82.74 ± 0.77
2	(40+20%) 60%	20%	Proposal	97.22 ± 0.53
			Naive Bayes	77.47 ± 0.90
			Log. Reg.	83.18 ± 0.81
3	(60+20%) 80%	20%	Proposal	97.74 ± 0.37
			Naive Bayes	77.79 ± 0.86
			Log. Reg.	85.03 ± 0.81

using the classification rule obtained from the training data. Next, the expert reviews the classification of this 20% and the classification rule is updated with the expert's information and also with the new prior distribution. Later, the second step considers more patients arriving to the system (20%). They are classified by using the previous classification rule. Then the expert supervises the results and both this information and the posterior distribution are aggregated to update the classification rule. The third and last step is performed after the two previous learning steps. It consists in classifying the last 20% of the data and evaluating the accuracy. Table 3 shows the results.

In the first step the confidence interval for the accuracy is 96.82 ± 0.63. By applying relevance feedback and including the posterior distribution as new prior for the second step, the method based on relevance feedback improves the results achieving 97.22 ± 0.53. Finally, in the third step, the information from the second one is used to learn and the results are better. In average, 97.74% of the patients have been correctly diagnosed. This accuracy is higher than the ones previously obtained by all the other methods without relevance feedback. The 95% confidence interval 97.74 ± 0.37 is the narrowest in the three steps. The more information is provided, the less variance values are obtained by using the proposed method. When comparing with logistic regression and Naive Bayes, the proposed approach provides better results in the three steps. Significance tests are used in the three steps. Non-parametric tests are used, because the homoscedasticity condition is not met. The variances provided by the three methods are statistically different ($p < 0.001$) in the three steps. Kruskal–Wallis' tests detected statistically significant differences among the percentages provided by the different methods ($p < 0.001$). All pairwise comparisons were statistically significant ($p < 0.001$). As in the case with no relevance feedback, the proposed method provided the largest accuracies, followed by the logistic regression one, and, finally, the Naive Bayes method provided the smallest accuracy.

Again, by using the proposed approach, there are no false negatives in all the iterations in the three steps. The specificities for the three steps are 0.9005, 0.914, and 0.9300, respectively. For the logistic regression, the sensitivity and specificity are: Step (1) 0.8527 and 0.7800, Step (2) 0.8680 and 0.7596, and Step (3) 0.8774 and 0.7935. For the Naive Bayes, the sensitivity and specificity are: Step (1) 0.7141 and 0.8696, Step (2) 0.7305 and 0.8631, and Step (3) 0.7379 and 0.8620.

The obtained results show the advantages of incorporating relevance feedback. With no relevance feedback, the method already achieves good results. However, by including the opinion of a physician through an iterative learning process, the classification method improves its performance, becoming a powerful tool in the computer-aided diagnosis context.

4. Conclusion

A novel classification method is proposed. It is based on a hybrid approach, exploiting the power of three different techniques: pairwise comparison, Bayesian regression and KNN. The method has the important advantage that it can be used for classification problems with both a large number of features and a few number of elements.

It is a flexible method, that can be applied automatically or in a supervised way. The performance of the method has been experimentally evaluated on two previously published datasets. By using the method in a fully automatic way, the results achieved by the proposed method clearly outperform the ones previously obtained. Additionally, in the second experiment the method improves the results every time it is executed, when it is applied with relevance feedback. This interactive methodology increases the quality of the results with the supervision of an expert and the readjustment of the prior distribution for the model parameters.

Although the approach is specially recommended to classify medical images, it is applicable to any type of biomedical objects described by vectors of features. The good results obtained in the experiments confirm its potential.

Conflict of interest

None.

Acknowledgments

The authors thank one anonymous referee for comments and suggestions which have improved the content and readability of this paper. This research has been partially funded by *Ministerio de Economí a y Competitividad, Spain* (Project MTM2011-28983-C03-02), *Junta de Extremadura, Spain* (Project GRU10110), and *European Union* (European Regional Development Funds).

REFERENCES

[1] R.O. Duda, P.E. Hart, H.G. Stork, Pattern Classification, Wiley-Interscience, New York, 2000.
[2] M.C. Bishop, Pattern Recognition and Machine Learning, Information Science and Statistics, Springer, New Yok, 2006.
[3] S. Theodoridis, K. Koutroumbas, Pattern Recognition, 4th ed., Academic Press, London, 2008.
[4] H. Handels, J. Ehrhardt, Medical image computing for computer-supported diagnostics and therapy. Advances and perspectives, Methods of Information in Medicine 48 (2009) 11–17.
[5] R. Smith-Bindman, D.L. Miglioretti, E.B. Larson, Rising use of diagnostic medical imaging in a large integrated health system, Health Affairs 27 (2008) 1491–1502.
[6] K. Doi, Computer-aided diagnosis in medical imaging: historical review, current status and future potential, Computerized Medical Imaging and Graphics 31 (2007) 198–211.
[7] J. Ramírez, J.M. Górriz, F. Segovia, R. Chaves, D. Salas-González, M. López, I. Álvarez, P. Padilla, Computer aided diagnosis system for the Alzheimer's disease based on partial least squares and random forest SPECT image classification, Neuroscience Letters 472 (2010) 99–103.
[8] M. Graña, M. Termenón, A. Savio, A. González-Pinto, J. Echeveste, J.M. Pérez, A. Besga, Computer aided diagnosis system for Alzheimer disease using brain diffusion tensor imaging features selected by Pearson's correlation, Neuroscience Letters 502 (2011) 225–229.
[9] D. Zhang, Y. Wang, L. Zhou, H. Yuan, D. Shen, The Alzheimer's disease neuroimaging initiative, multimodal classification of Alzheimer's disease and mild cognitive impairment, NeuroImage 55 (2011) 856–867.
[10] F.J. Martínez-Murcia, J.M. Górriz, J. Ramírez, C.G. Puntonet, D. Salas-González, Computer aided diagnosis tool for Alzheimer's disease based on Mann–Whitney–Wilcoxon U-test, Expert Systems with Applications 39 (2012) 9676–9685.
[11] S. Kasai, F. Li, J. Shiraishi, Q. Li, K. Doi, Computerized detection of vertebral compression fractures on lateral chest radiographs: preliminary results of a tool for early detection of osteoporosis, Medical Physics 33 (2006) 4664–4674.
[12] S. Kasai, F. Li, J. Shiraishi, K. Doi, Usefulness of computer-aided diagnosis schemes for vertebral fractures and lung nodules on chest radiographs, American Journal of Roentgenology 191 (2008) 260–265.
[13] D. Voigt, M. Dollinger, A. Yanga, U. Eysholdta, J. Lohscheller, Automatic diagnosis of vocal fold paresis by employing phonovibrogram features and machine learning methods, Computer Methods and Programs in Biomedicine 99 (2010) 275–288.
[14] A.E. Zadeh, A. Khazaee, V. Ranaee, Classification of the electrocardiogram signals using supervised classifiers and efficient features, Computer Methods and Programs in Biomedicine 99 (2010) 179–194.
[15] J.M. Górriz, E. Lang, J. Ramírez, Recent advances in biomedical signal processing, Bentham Science Publishers, 2011.
[16] Baghai-Ravary, Ladan, W. Steve, Beet, Automatic speech signal analysis for clinical diagnosis and assessment of speech disorders, Springer, New York, Heidelberg, Dordrecht, London, 2013.
[17] A. Tsanas, M.A. Little, P.E. McSharry, J. Spielman, L.O. Ramig, Novel speech signal processing algorithms for high-accuracy classification of Parkinson's disease, IEEE Transactions on Biomedical Engineering 59 (2012) 1264–1271.
[18] I.A. Illán, J.M. Górriz, J. Ramírez, D. Salas-González, M. López, F. Segovia, P. Padilla, C.G. Puntonet, Projecting independent components of SPECT images for computer aided diagnosis of Alzheimer's disease, Pattern Recognition Letters 31 (2010) 1342–1347.
[19] B. Song, G. Zhang, H. Wang, W. Zhu, Z. Liang, A dimension reduction strategy for improving the efficiency of computer-aided detection for CT colonography, in: Proceedings of SPIE Medical Imaging, 2013, p. 86702A.
[20] R.B. Reilly, R. Moran, P. Lacy, Voice pathology assessment based on a dialogue system and speech analysis, in: Proceedings of the American Association of Artificial Intelligence Fall Symposium on Dialogue Systems for Health Communication, 2004, pp. 104–109.
[21] V. van Ravesteijn, C. van Wijk, F. Vos, R. Truyen, J. Peters, J. Stoker, L. van Vliet, Computer aided detection of polyps in CT colonography using logistic regression, IEEE Transactions on Medical Imaging 29 (2010) 120–131.
[22] R. Chaves, J. Ramírez, J.M. Górriz, M. López, D. Salas-González, I. Álvarez, F. Segovia, SVM-based computer aided diagnosis of the Alzheimer's disease using t-test NMSE feature selection with feature correlation weighting, Neuroscience Letters 461 (2009) 293–297.
[23] O.A. Debats, N. Karssemeijer, J.O. Barentsz, H. Huisman, Automated classification of lymph nodes in USPIO-enhanced MR images: a comparison of three segmentation methods, in: Medical Imaging 2010: Computer-Aided Diagnosis, volume 11 of Proceedings of SPIE, 2010, pp. 7624–7625.
[24] K. Murphy, B. van Ginneken, A.M. Schilham, B.J. de Hoop, H.A. Gietema, M. Prokop, A large-scale evaluation of automatic pulmonary nodule detection in chest CT using local image features and k-nearest-neighbour classification, Medical Image Analysis 13 (2009) 757–770.
[25] S. Joo, Y.S. Yang, W.K. Moon, H.C. Kim, Computer-aided diagnosis of solid breast nodules: Use of an artificial neural network based on multiple sonographic features, IEEE Transactions on Medical Imaging 23 (2004) 1292–1300.
[26] M. López, J. Ramírez, J.M. Górriz, D. Salas-González, I. Álvarez, F. Segovia, C.G. Puntonet, Automatic tool for Alzheimer's disease diagnosis using PCA and Bayesian classification rules, Electronics Letters 45 (2009) 389–391.
[27] S.H. Raza, Y. Sharma, Q. Chaudry, A.N. Young, M.D. Wang, Automated classification of renal cell carcinoma subtypes using scale invariant feature transform., in: Engineering in Medicine and Biology Society, 2009, pp. 6687–6690.
[28] J. Hayward, S.A. Álvarez, C. Ruiz, M. Sullivan, J. Tseng, G. Whalen, Machine learning of clinical performance in a pancreatic cancer database, Artificial Intelligence in Medicine 49 (2010) 187–195.
[29] L. Breiman, Bagging predictors, Machine Learning 24 (1996) 123–140.
[30] Y. Freund, R.E. Schapire, A decision-theoretic generalization of on-line learning and an application to boosting, Journal of Computer and System Sciences 55 (1997) 119–139.
[31] P. Mehrotra, J. Chatterjee, C. Chakraborty, B. Ghoshdastidar, S. Ghoshdastidar, Automated screening of polycystic ovary syndrome using machine learning techniques, in: India Conference (INDICON), 2011 Annual IEEE, 2011, pp. 1–5.
[32] B. Mwangi, K.P. Ebmeier, K. Matthews, J.D. Steele, Multi-centre diagnostic classification of individual structural neuroimaging scans from patients with major depressive disorder, Brain 135 (2012) 1508–1521.
[33] J.P. Arias-Nicolás, F. Calle, I.M. Horrillo, J.R. Martín, QATRIS IManager, a Bayesian Image Search Engine, 2008, International Society for Business and Industrial Statistics 2008 Book of abstracts.

[34] L.L. Thurstone, A law of comparative judgements, Psychological Review 34 (1927) 273–286.
[35] J.P. Arias-Nicolás, C.J. Pérez, J.R. Martín, A logistic regression-based pairwise comparison method to aggregate preferences, Group Decision and Negotiation 17 (2008) 237–247.
[36] M.L. Durán, P.G. Rodríguez, J.P. Arias-Nicolás, J. Martín, C. Disdier, A perceptual similarity method by pairwise comparison in a medical image case, Machine Vision and Applications 21 (2010) 865–877.
[37] J.H. Albert, S. Chib, Bayesian analysis of binary and polychotomous response data, Journal of the American Statistical Association 88 (1993) 669–679.
[38] D.J. Spiegelhalter, N.G. Best, B.P. Carlin, A. van der Linde, Bayesian measures of model complexity and fit, Journal of the Royal Statistical Society 64 (2002) 583–639.
[39] A. Zellner, P. Rossi, Bayesian analysis of dichotomous quantal response models, Journal of Econometrics 25 (1994) 365–393.
[40] E. Fix, J.L. Hodges, Discriminatory analysis, nonparametric discrimination consistency properties, USAF School of Aviation Medicine, 1951, Technical Report 4.
[41] S. Tan, Neighbor-weighted k-nearest neighbor for unbalanced text corpus, Expert System Applications 28 (2005) 667–671.
[42] F. Rosenblatt, The perceptron: a probabilistic model for information storage and organization in the brain, Psychological Review 65 (1958) 386–408.
[43] A. Frank, A. Asuncion, UCI machine learning repository (2010).
[44] J. Estrela da Silva, J.P. Marques de Sá, J. Jossinet, Classification of breast tissue by electrical impedance spectroscopy, Medical and Biological Engineering and Computing 38 (2000) 26–30.
[45] J. Jossinet, Elementary electrodynamics, Technology and Health Care 16 (2008) 465–474.
[46] I. Streitner, M. Goldhofer, S. Cho, H. Thielecke, R. Kinscherf, F. Streitner, J. Metz, K. Haase, M. Borggrefe, T. Suselbeck, Electric impedance spectroscopy of human atherosclerotic lesions, Atherosclerosis 206 (2009) 464–468.
[47] P. Aberg, U. Birgersson, P. Elsner, P. Mohr, S. Ollmar, Electrical impedance spectroscopy and the diagnostic accuracy for malignant melanoma, Experimental Dermatology 20 (2011) 648–652.
[48] R.J. Halter, A. Schned, J. Heaney, A. Hartov, S. Schutz, K.D. Paulsen, Electrical impedance spectroscopy of benign and malignant prostatic tissues, The Journal of Urology 179 (2008) 1580–1586.
[49] J. Jossinet, The impedivity of freshly excised human breast tissue, Physiology Measures 19 (1998) 61–75.
[50] D. Lederman, B. Zheng, X. Wang, X. Wang, D. Gur, Improving breast cancer risk stratification using resonance-frequency electrical impedance spectroscopy through fusion of multiple classifiers, Annals of Biomedical Engineering 39 (2011) 931–945.
[51] J. Jossinet, B. Lavandier, The discrimination of excised cancerous breast tissue samples using impedance spectroscopy, Bioelectrochemistry and Bioenergetics 45 (1998) 161–167.
[52] P.H. Swain, H. Hauska, The decision tree classifier: design and potential, IEEE Transaction on Geoscience and Remote Sensing GE-15 (1977) 142–147.
[53] K. Suzuki, Machine learning in computer-aided diagnosis: medical imaging intelligence and analysis, Medical Information Science Reference, IGI Global, 2012.
[54] A.R. da Rocha Neto, R. Sousa, G. de, A. Barreto, J.S. Cardoso, Diagnostic of pathology on the vertebral column with embedded reject option, in: Proceedings of the 5th Iberian conference on pattern recognition and image analysis, Springer-Verlag, 2011, pp. 588–595.
[55] E. Berthonnaud, J. Dimnet, P. Roussouly, H. Labelle, Analysis of the sagittal balance of the spine and pelvis using shape and orientation parameters, Journal of Spinal Disorders and Techniques 18 (2005) 40–47.
[56] R. Sousa, B. Mora, J.S. Cardoso, An ordinal data method for the classification with reject option, in: International Conference on Machine Learning and Applications ICMLA'09, pp. 746–750.
[57] A.R.R. Neto, G.A. Barreto, On the application of ensembles of classifiers to the diagnosis of pathologies of the vertebral column: a comparative analysis, IEEE Latin America Transactions 7 (2009) 487–496.
[58] J.P.A. Ioannidis, R. Tarone, J.K. McLaughlin, The false-positive to false-negative ratio in epidemiologic studies, Epidemiology 22 (2011) 450–456.