Artificial Intelligence in Medicine xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

Artificial Intelligence in Medicine

journal homepage: www.elsevier.com/locate/aiim

Improving the Mann–Whitney statistical test for feature selection: An approach in breast cancer diagnosis on mammography

Noel Pérez Pérez[a,*], Miguel A. Guevara López[b,a], Augusto Silva[b], Isabel Ramos[c]

[a] *Institute of Mechanical Engineering and Industrial Management (INEGI), Campus da FEUP, Rua Dr. Roberto Frias, 400, 4200-465 Porto, Portugal*
[b] *Institute of Electronics and Telematics Engineering of Aveiro, Campus Universitário de Santiago, 3810-193 Aveiro, Portugal*
[c] *Faculty of Medicine – Centro Hospitalar São Joao, Al. Prof. Hernâni Monteiro, 4200-319 Porto, Portugal*

ARTICLE INFO

Article history:
Received 24 January 2014
Received in revised form
21 November 2014
Accepted 4 December 2014

Keywords:
Feature selection methods
Mann–Whitney *U*-test
uFilter method
Machine learning algorithms
Redundancy analysis
Breast cancer CADx

ABSTRACT

Objective: This work addresses the theoretical description and experimental evaluation of a new feature selection method (named uFilter). The uFilter improves the Mann–Whitney *U*-test for reducing dimensionality and ranking features in binary classification problems. Also, it presented a practical uFilter application on breast cancer computer-aided diagnosis (CADx).
Materials and methods: A total of 720 datasets (ranked subsets of features) were formed by the application of the chi-square (CHI2) discretization, information-gain (IG), one-rule (1Rule), Relief, uFilter and its theoretical basis method (named *U*-test). Each produced dataset was used for training feed-forward backpropagation neural network, support vector machine, linear discriminant analysis and naive Bayes machine learning algorithms to produce classification scores for further statistical comparisons.
Results: A head-to-head comparison based on the mean of area under receiver operating characteristics curve scores against the *U*-test method showed that the uFilter method significantly outperformed the *U*-test method for almost all classification schemes ($p < 0.05$); it was superior in 50%; tied in a 37.5% and lost in a 12.5% of the 24 comparative scenarios. Also, the performance of the uFilter method, when compared with CHI2 discretization, IG, 1Rule and Relief methods, was superior or at least statistically similar on the explored datasets while requiring less number of features.
Conclusions: The experimental results indicated that uFilter method statistically outperformed the *U*-test method and it demonstrated similar, but not superior, performance than traditional feature selection methods (CHI2 discretization, IG, 1Rule and Relief). The uFilter method revealed competitive and appealing cost-effectiveness results on selecting relevant features, as a support tool for breast cancer CADx methods especially in unbalanced datasets contexts. Finally, the redundancy analysis as a complementary step to the uFilter method provided us an effective way for finding optimal subsets of features without decreasing the classification performances.

1. Introduction

Devijver and Kittler define Feature Selection as the problem of "*extracting from the raw data the information which is most relevant for classification purposes, in the sense of minimizing the within-class pattern variability while enhancing the between-class pattern variability*" [1]. Guyon and Elisseeff consider that "*feature selection addresses the problem of finding the most compact and informative set of features, to improve the efficiency or data storage and processing*" [2].

During the last decade parallel efforts from researchers in statistics, machine learning, and knowledge discovery have been focused on the problem of feature selection and its influence in machine learning classifiers. The recent advances made in both sensing technologies and machine learning techniques make it possible to design recognition systems, which are capable of performing tasks that could not be performed in the past [2]. Feature selection lies at the center of these advances with applications in the pharmaceutical industry [3,4], oil industry [5,6], speech recognition [7,8], pattern recognition [9,10], biotechnology [11,12] and many other

[*] Corresponding author. Tel.: +351 229578710; fax: +351 229537352.
E-mail address: nperez@inegi.up.pt (N.P. Pérez).

http://dx.doi.org/10.1016/j.artmed.2014.12.004
0933-3657/© 2014 Elsevier B.V. All rights reserved.

emerging fields with significant impact in health systems for cancer detection [13–17].

In contrast to other dimensionality reduction techniques like those based on projection (e.g. principal component analysis) or compression (e.g. using information theory), feature selection techniques do not alter the original representation of the variables, but merely select a subset of them. Thus, they preserve the original semantics of the variables, hence, offering the advantage of interpretation by a domain expert [18].

There are many potential benefits of feature selection: facilitating data visualization and data understanding, reducing the measurement and storage requirements, reducing training and utilization times, defining the curse of dimensionality to improve the predictions performance [9]. The objectives are related: to avoid overfitting and improve model performance; to provide faster and more cost-effective models, and to gain a deeper insight into the underlying processes that generated the data [18]. Although these benefits, the problem of finding or ranking relevant features is still a challenging task.

Regarding their classification, feature selection techniques can be structured into three paradigms, depending on how they combine the feature selection search with the construction of the classification model: wrappers, embedded and filters methods (univariate and multivariate). Wrappers utilize machine learning classifiers as a black box to score subsets of features according to their predictive power. Embedded methods perform feature selection in the process of training and are usually specific to given machine learning classifiers. Filters methods (considered the earliest approaches) use heuristics based on general characteristics of the data rather than machine learning classifiers to evaluate the merit of features [2,9]. Therefore, filter methods in general present lower algorithm complexity and are much faster than wrapper or embedded methods [2,9].

Univariate filter methods, such as chi-square (CHI2) discretization [19], *t*-test [20], information gain (IG) [21] and gain ratio [22], present two main disadvantages: (1) ignoring the dependencies among features and (2) assuming a given distribution (Gaussian in most cases) from which the samples (observations) have been collected. In addition, to assume a Gaussian distribution includes the difficulties to validate distributional assumptions because of small sample sizes. On the other hand, multivariate filters methods such as: correlation based-feature selection [20,23], Markov blanket filter [24], fast correlation based-feature selection [10], ReliefF [25,26] overcome the problem of ignoring feature dependencies introducing redundancy analysis (models feature dependencies) at some degree, but the improvements are not always significant: domains with large numbers of input variables suffer from the curse of dimensionality and multivariate methods may overfit the data. Also, they are slower and less scalable than univariate methods [2,9].

To overcome these inconveniences we developed the uFilter method. uFilter is an innovative feature selection method for ranking relevant features that assess the relevance of features by computing the separability between class-data distribution of each feature. We address a theoretical description and experimental evaluation of the uFilter method, and it is described a formal framework for understanding the proposed algorithm, which is supported on the statistical model/theory of the non-parametric Mann–Whitney *U*-test [27]. A software prototype implementation of the uFilter method using these theoretical intuitions is also presented. The uFilter is an univariate filter method that solves some difficulties remaining on previous methods, such as: (1) it is effective in ranking relevant features independently of the samples sizes (tolerant to unbalanced training data); (2) it does not need any type of data normalization; and (3) the most important, it presents a low risk of data overfitting and does not incur the high computational cost of conducting a search through the space of feature subsets as in the wrapper or embedded methods.

The remainder of the paper is organized as follows: Section 2 describes in detail the developed uFilter method as well as the experimental methodology for its evaluation in breast cancer databases; Section 3 presents and discusses the experimental results obtained both from the head-to-head statistical comparison between the proposed method and the theoretical basis (named *U*-test) method and from the global comparison with other well-known feature selection methods. Finally, in Section 4 we outline the principal achievements of the work.

2. Materials and methods

2.1. Databases

This work considered two public databases: the Breast Cancer Digital Repository (BCDR) and the Digital Database for Screening Mammography (DDSM). The BCDR is the first wide-ranging annotated Portuguese breast cancer repository, with anonymous cases from medical historical archives supplied by Faculty of Medicine – *Centro Hospitalar de São João* at University of Porto, Portugal [28,29]. For convenience, the DDSM images used in this study were obtained from the Image Retrieval in Medical Applications (IRMA) project (courtesy of TM Deserno, Dept. of Medical Informatics, RWTH Aachen, Germany) where the original LJPEG images of DDSM were converted to 16 bits portable network graphics format [30,31].

BCDR is composed of 1734 patient cases with mammography and ultrasound images, clinical history, lesion segmentation and selected pre-computed image-based descriptors; each case may have one or more segmented (outlined) region of interest (ROI) associated to a pathological lesion, typically in mediolateral oblique (MLO) and craniocaudal (CC) images of the same breast [28,29]. We used the dataset BCDR-F01 available online at the BCDR website (http://bcdr.inegi.up.pt), which is composed by 190 patient cases, biopsy proven.

DDSM database is composed by 2620 patient cases divided into three categories: normal cases (12 volumes), cancer cases (15 volumes) and benign cases (14 volumes); each case may have one or more associated pathological lesion segmentations, usually in MLO and CC image views of the same breast [30,31]. Due to the wide range of information, we considered a dataset formed with 582 patient cases representing two volumes of cancer and benign cases (random selected).

2.2. Considered features and datasets creation

Each instance of the datasets (above mentioned) is composed by a set of 23 image-based descriptors extracted both for the BCDR and DDSM databases. This rather extensive feature list builds upon the radiologists experience and previously reported feature lists embedded in breast cancer computer-aided diagnosis (CADx) systems. Selected descriptors included intensity statistics, shape and texture features computed from segmented pathological lesions in both MLO and CC mammography views. The intensity statistics and shape descriptors were selected according to the radiologists experience (similar to the clinician procedure) and the American College of Radiology (BIRADS-Mammography atlas) [32], which described in detail how to detect/classify pathological lesions. Additionally, texture descriptors were the Halarick's descriptors extracted from the gray-level co-occurrence matrices [33]. An overview of the mathematical formulation is presented below:

- Skewness:

$$f_1 = \frac{(1/n)\sum_{i=1}^{n}(x_i - \bar{x})^3}{\left(\sqrt{(1/n)\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)^3}$$

  with x_i being the ith-value and $\bar{x}$ the sample mean.
- Kurtosis:

$$f_2 = \frac{(1/n)\sum_{i=1}^{n}(x_i - \bar{x})^4}{\left((1/n)\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^2} - 3$$

  with x_i being the ith-value and $\bar{x}$ the sample mean.
- Circularity:

$$f_3 = 4\pi\frac{\text{area}}{\text{perimeter}^2}$$

- Perimeter:

$$f_4 = \text{length}(E)$$

  with $E \subset O$ being the edge pixels.
- Elongation:

$$f_5 = \frac{m}{M}$$

  with m being the minor axis and M the major axis of the ellipse that has the same normalized second central moments as the region surrounded by the contour.
- Standard deviation:

$$f_6 = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

  with x_i being the gray level intensity of the ith pixel and $\bar{x}$ the mean of intensity.
- Roughness:

$$f_7 = \frac{(\text{perimenter})^2}{4\pi * \text{area}}$$

- Minimum (f_8) and maximum (f_9): the minimum and maximum intensity value in the region surrounded by the contour.
- Shape:

$$f_{10} = \frac{\text{perimeter} * \text{elongation}}{8 * \text{area}}$$

- X centroid:

$$f_{11} = \frac{\min(x) + \max(x)}{2}$$

  with x being the set of X coordinates of the object's contour.
- Entropy:

$$f_{12} = -\sum_{i=1}^{L}\sum_{j=1}^{L} p(i,j)\log(p(i,j))$$

  with $p(i,j)$ being the probability of pixels with gray-level i occur together to pixels with gray-level j.
- X center mass (f_{13}) and Y center mass (f_{14}): normalized X and Y coordinates of the center of mass of O
- Angular second moment:

$$f_{15} = \sum_{i=1}^{L}\sum_{j=1}^{L} p(i,j)^2$$

  with L being the number of gray-levels, and p being the gray-level co-occurrence matrix and, thus, $p(i,j)$ is the probability of pixels with gray-level i occur together to pixels with gray-level j.
- Median:

$$f_{16} = \begin{cases} \text{MED} = \dfrac{n+1}{2}, & \text{if length}(X)\text{ is odd} \\[2ex] \text{MED} = \dfrac{X\left(\frac{n}{2}\right) + X\left(\frac{n}{2}+1\right)}{2}, & \text{if length}(X)\text{ is even} \end{cases}$$

  with X being the set of intensities.
- Contrast:

$$f_{17} = \sum_{i}\sum_{j}(i-j)^2 p(i,j)$$

  with $p(i,j)$ being the probability of pixels with gray-level i occur together to pixels with gray-level j.
- Correlation:

$$f_{18} = \frac{\sum_i\sum_j (ij)p(i,j) - \mu_x\mu_y}{\sigma_x\sigma_y}$$

  with μ_x, μ_y, σ_x and σ_y being the means and standard deviations of the marginal distribution associated with $p(i,j)$.
- Mean:

$$f_{19} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

  with n being the number of pixels inside the region delimited by the contour and x_i being the gray level intensity of the ith pixel inside the contour.
- Inverse difference moment:

$$f_{20} = \sum_{i}\sum_{j}\frac{1}{1+(i-j)^2}p(i,j)$$

  with $p(i,j)$ being the probability of pixels with gray-level i occur together to pixels with gray-level j.
- Area:

$$f_{21} = |O|$$

  with O being the set of pixels that belong to the segmented lesion.
- Y centroid:

$$f_{22} = \frac{\min(Y) + \max(Y)}{2}$$

- with Y being the set of Y coordinates of the object's contour.
- Statistical mode (f_{23}): most frequent intensity value in a segmented ROI (lesion).

Please cite this article in press as: Pérez NP, et al. Improving the Mann–Whitney statistical test for feature selection: An approach in breast cancer diagnosis on mammography. Artif Intell Med (2014), http://dx.doi.org/10.1016/j.artmed.2014.12.004

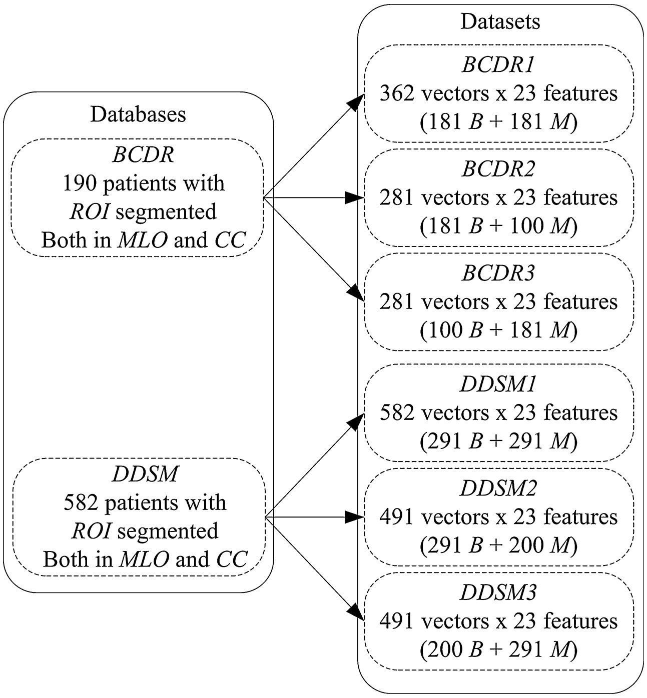

Fig. 1. Datasets creation; *B* and *M* represent benign and malignant class instances.

Taking as start point the two initially formed datasets from BCDR and DDSM, we created six new datasets (three from BCDR and three from DDSM respectively) representing three different configurations: (1) two balanced datasets (same quantity of benign and malignant instances), (2) two unbalanced datasets containing more benign than malignant instances and (3) two unbalanced datasets holding more malignant than benign instances. From the BCDR, we created the BCDR1 dataset comprising 362 features vectors, and the BCDR2 and BCDR3 datasets including a total of 281 features vectors (each one). Besides, from the DDSM database, we formed the DDSM1 dataset holding 582 features vectors, and the DDSM2 and DDSM3 datasets involving a total of 491 features vectors respectively. Fig. 1 shows a detail description of created datasets.

2.3. Feature selection methods

Many types of extracted features (e.g. intensity statistics, shape and texture) from mammograms have been used to form subsets of features with significant information about different lesions [13,15,17,34]. However, selecting the most appropriate subset of features is still a very difficult task; usually a satisfactory instead of the optimal feature subset is searched.

Dash and Liu in Ref. [35] stated that an optimal subset is always relative to a certain evaluation function. It is mean that an optimal subset chosen using one evaluation function may not be the same as that which uses another evaluation function. An example of this is the work of Ghazavi and Liao [14], which used different evaluation functions for feature selection purposes, including three modalities of mutual correlation, two variants of Welch *t*-statistic, two variants of Fisher correlation, the independently consistent expression discriminator, and two distance scores. These functions were evaluated on two binary class medical datasets available at UCI repository: the Wisconsin breast cancer and Pima Indians diabetes, and on a particular industrial dataset: the welding Flaw. The reported results highlighted the mutual correlation method as the best feature selector for the Wisconsin breast cancer and welding flaw datasets respectively. Meanwhile, the best result in the Pima Indians diabetes dataset was achieved by either one of the four statistical criteria.

Usually, an evaluation function tries to measure the discriminating ability of a feature or a subset to distinguish the different class' labels. Thus, the use of different evaluation function provides important information about the nature of each feature (respect to the class) in the features space. We selected different methods with different evaluation function, all of them derived from the filter paradigm (independent of classifiers):

- CHI2 discretization: this method consists on a justified heuristic for supervised discretization [19]. Numerical features are initially sorted by placing each observed value into its own interval. Then the chi-square statistic is used to determine whether the relative frequencies of the classes in adjacent intervals are similar enough to justify merging. The extent of the merging process is controlled by an automatically set chi-square threshold. The threshold is determined through attempting to maintain the fidelity of the original data.
- IG method: the IG measurement normalized with the symmetrical uncertainty coefficient [21] is a symmetrical measure in which the amount of information gained about *Y* after observing *X* is equal to the amount of information gained about *X* after observing *Y* (a measure of feature – feature intercorrelation). This model is used to estimate the value of an attribute *Y* for a novel sample (drawn from the same distribution as the training data) and compensates for information gain bias toward attributes with more values.
- 1Rule: this method estimates the predictive accuracy of individual features building rules based on a single feature (can be thought of as single level decision trees) [36]. As it is used training and test datasets, it is possible to calculate a classification accuracy for each rule and hence each feature. Then, from the classification scores, a ranked list of features is obtained. Experiments with choosing a selected number of the highest ranked features and using them with common machine learning algorithms showed that, on average, the top three or more features are as accurate as using the original set. This approach is unusual due to the fact that no search is conducted.
- Relief: this method uses instance based learning to assign a relevance weight to each feature [25]. Each feature weight reflects its ability to distinguish among the class values. The feature weight is updated according to how well its values distinguish the sampled instance from its nearest hit (instance of the same class) and nearest miss (instance of opposite class). The feature will receive a high weight if it differentiates between instances from different classes and has the same value for instances of the same class. For nominal features it is defined as either 1 (the values are different) or 0 (the values are the same), while for numeric features the difference is the actual difference normalized to the interval [0.1].

2.4. Machine learning classifiers

The discrimination of two different samples is a supervised learning problem, which is defined as the prediction of the value of a function for any valid input after training a learner using examples of input and target output pairs [27]. For the problem at hand, the function has only two discrete values: benign or malignant. Hence the problem of discriminating benign and malignant lesions can be modeled as a two-class classification problem.

Among a wide variety of machine learning classifiers that have been applied in mammography-based CADx systems to solve the problem of breast cancer classification; artificial neural networks (ANN) [37–43], support vector machines (SVM) [40,44–46] and linear discriminant analysis (LDA) [16,41,42,47–50] seem to be the

Please cite this article in press as: Pérez NP, et al. Improving the Mann–Whitney statistical test for feature selection: An approach in breast cancer diagnosis on mammography. Artif Intell Med (2014), http://dx.doi.org/10.1016/j.artmed.2014.12.004

most commonly used type of classifiers. Other less used with high performance in breast cancer classification are the Naive Bayes (NB) classifier [37,51–53] and the fuzzy modeling methods [54–57]. However, the latter is very expensive in terms of CPU time consuming, because they are mainly based on rules. An example is the work of Ghazavi and Liao [14] where three fuzzy modeling methods for breast cancer classification were used, achieving a satisfactory AUC value (0.9587) when using the fuzzy k-nearest neighbor algorithm in the Wisconsin breast cancer dataset, but the CPU time consumed was high.

We used the ANN, SVM, LDA and NB classifiers implemented and available on the Weka version 3.6 [58]. For all classifiers with the exception of the NB (which is parameterless), a 10-fold cross validation method [59] was applied on the training set for optimizing classifiers' parameters. A brief description and configuration of employed machine learning classifiers is given here:

- The feed forward back-propagation (FFBP) neural network is a particular model of ANN, which provides a nonlinear mapping between its input and output according to the back-propagation error learning algorithm [60]. We used this classifier with the following parameters: neurons on hidden layers were determined according to the equation (attributes + number of classes)/2; one output layer associated with the binary classification (benign or malignant); the sigmoid function was used as transfer function on all layers and the number of iterations (epochs) were optimized in the range of 100–1000 epochs (with an interval increment of 100 units).
- SVMs are based on the definition of an optimal hyperplane, which linearly separates the training data. In comparison with other classification methods, a SVM aims to minimize the empirical risk and maximize the distances (geometric margin) of the data points from the corresponding linear decision boundary [60]. The SVM classifier was used with the following settings: the regularization parameter C (cost) was optimized in the range of 10^{-3}–10^{3} and the kernel type was based on a linear function, which provided better results respect to others kernel such as: radial basis, polynomial and sigmoid function (from our experimental experience).
- LDA is a traditional method for classification [61]. The basic idea is to try to find an optimal projection (decision boundaries optimized by the error criterion), which can maximize the distances between samples from different classes and minimize the distances between samples from the same class. We used LDA for binary classification, thus, observations were classified by the following linear function:

  $$g_i(x) = W_i^T * x - c_i, \quad 1 \le i \le 2$$

  where W_i^T is the transpose of a coefficient vector, x is a feature vector and c_i is a constant as the threshold. The values of W_i^T and c_i are determined through the analysis of a training set. Once these values are determined, they can be used to classify the new observations (smallest $g_i(x)$ is preferred).
- The NB classifier is based on probabilistic models with strong (Naive) independence assumptions, which assumes a class variable depending on the set of input features [62]. This classifier can be trained based on the relative frequencies shown in the training set to get an estimation of the class priors and feature probability distributions. For a test sample, the decision rule will be picking the most probable hypothesis, which is known as the maximum a posteriori decision rule.

2.5. The uFilter method

We considered developing the new uFilter feature selection method based on the Mann–Whitney *U*-test [27], in a first approach, to be applied in binary classification problems. The uFilter algorithm is framed in the univariate filter paradigm since it requires only the computation of *n* scores and sorting them. Therefore, its time execution (faster) and complexity (lower) are beneficial when is compared to wrapper or embedded methods.

The Mann–Whitney *U*-test is a non-parametric method used to test whether two independent samples of observations are drawn from the same or identical distributions. *U*-test is based on the idea that the particular pattern exhibited when *m* number of *X* random variables and *n* number of *Y* random variables are arranged together in increasing order of magnitude provides information about the relationship between their parent populations [27]. The Mann–Whitney test criterion is based on the magnitude of the *Y*'s in relation to the *X*'s (e.g. the position of *Y*'s in the combined ordered sequence). A sample pattern of arrangement where most of the *Y*'s are greater than most of the *X*'s or vice versa would be evidence against random mixing. This would tend to discredit the null hypothesis of identical distribution [63]. An advantage of using this test is that the two samples under consideration may not necessarily have the same number of observations. However, samples with at least 25 instances would be more desirable for statistical analysis, since data distribution follows the normal distribution.

For better understanding the theoretical description, we considered a binary class problem (benign and malignant classes) with more than 25 instances per class, thus:

Let $F = \{f_1, f_2, \ldots, f_t\}$ a set of features with size t, and let $f_i = \{I_{c,1}, I_{c,2}, \ldots, I_{c,n}\}$ an ordered set of instances (in ascending way) with size n belonging to the ith-feature under analysis, where $I_{c,j}$ represents the value of the feature f_i for an individual instance j, and c denotes the class value: Benign (B) and Malignant (M). Then, the uFilter performs a tie analysis in the f_i sequence according to the rule: if there are tie elements, their positions are updated by the resultant value of averaging the positions of tied elements; the output sequence is saved in f_i'. Consequently, summation of benign (S_B) and malignant (S_M) instance positions in the f_i' sequence was defined by:

$$S_B = \sum_{j=1}^{n_B} f_j' \tag{1}$$

$$S_M = \sum_{j=1}^{n_M} f_j' \tag{2}$$

where n_B and n_M are the totals of benign and malignant instances respectively. Thus, *u*-values (according to the Mann–Whitney *U*-test [27]) for each sample are computed as:

$$uB = n_B n_M + \frac{n_B(n_B+1)}{2} - S_B \tag{3}$$

$$uM = n_B n_M + \frac{n_M(n_M+1)}{2} - S_M \tag{4}$$

As the sample size exceeds 25 instances per class, the original Mann–Whitney *U*-test [27] selected the minimum between both computed *u*-values (from Eqs. (3) to (4)) for the calculation of the *Z*-indicator (see Eqs. (5) or (6)). In this case, only one *Z*-indicator will be analyzed to accept or reject the null hypothesis at a given level of significance ($\alpha = 0.05$).

In contrast with the original *U*-test, the proposed uFilter method computes both *Z*-indicators (one by each class) in the following way:

$$Z_B = \frac{uB - \overline{u}}{\sigma_u} \tag{5}$$

Please cite this article in press as: Pérez NP, et al. Improving the Mann–Whitney statistical test for feature selection: An approach in breast cancer diagnosis on mammography. Artif Intell Med (2014), http://dx.doi.org/10.1016/j.artmed.2014.12.004

$$Z_M = \frac{uM - \bar{u}}{\sigma_u} \tag{6}$$

where $\bar{u}$ is the mean of the sample and the standard deviation is defined as:

$$\sigma_u = \sqrt{\frac{n_B n_M}{n(n-1)}\left(\frac{n^3-n}{12}\right) - \sum_{i=1}^{k}\frac{l_i^3 - l_i}{12}}$$

where k denotes the total of range having tied elements in f_i sequence and l_i means the quantity of elements within each kth-range. Finally, the score/weight of the feature f_i is calculated as the absolute value of the numerical difference between Z scores (see Eq. (7)).

$$w_i = |Z_B - Z_M| \tag{7}$$

The uFilter algorithm is applied to the whole feature space and the output of the algorithm is the ranking of features established by sorting in descendant way the random sequence of weights (w). In this approach, higher values in Eq. (7) are preferred, because it means the feature has better separability of class-data distributions and therefore higher discrimination power. Otherwise, the class-data distributions is overlapping and finding the decision boundary for future classifications becomes more difficult. Algorithm 1 summarizes the uFilter steps.

2.6. Experimental methodology

This section outlines the experimental evaluation of the proposed uFilter method when compared to four well known (classical) feature selection methods on breast cancer diagnosis. Since this research is a multistep modeling procedure, the application of the k-fold cross validation method [59] to the entire sequence of modeling steps guarantees reliable results [64].

In particular, we applied ten times the 10-fold cross validation method before to establish a ranking of features (to avoid giving an unfair advantage to predictors) and classification steps respectively (to prevent overfitting of classifiers to the training set [59]) (see Fig. 2). Thus, no samples appear simultaneously in training and test (disjoint test partitions). In this way, individual classifiers will be trained on different training sets, leading to different representations of the input space. Testing on these different input space representations leads to diversity in the resultant classifications for individual samples.

The overall procedure for the uFilter evaluation involves five main steps:

- Applying the classical Mann–Whitney U-test (U-test), the new proposed uFilter method and four well known feature selection methods: CHI2 discretization [19], IG [21], 1Rule [36] and Relief [25] to the six previously formed breast cancer datasets (see Fig. 2 step 2);
- Creating several ranked subset of features using increasing quantities of features. The top N features of each ranking (resultant from the previous step) were used for feeding different classifiers, with N varying from 5 to the total number of features of the dataset, with increments of 5 (see Fig. 2 step 3).
- Classifying the generated ranked subset of features using FFBP neural network [60], SVM [60], LDA [61] and NB [62] classifiers for a comparative analysis of AUC scores. All comparisons were using the Wilcoxon statistical test [65,66] to assess the meaningfulness of differences between classification schemes (see Fig. 2 step 3);
- Selecting the best classification scheme on datasets (BCDR1, BCDR2, BCDR3, DDSM1, DDSM2 and DDSM3), and thus the best subset of features.

In the last step of the experiment, we determined the feature relevance analysis using a two-step procedure involving (1) selecting the best subset of features for each dataset, and (2) performing a redundancy analysis based on the Pearson correlation [67], to determine and eliminate redundant features from relevant ones, and thus to produce the final subset of features.

In contrast to the work of Ghazavi and Liao [14], we decided to employ the correlation analysis as a complementary step to the uFilter procedure, instead to an evaluation function for features selection, because in real domains many features have high correlations and thus many are (weakly) relevant and should not be removed [68]. Also, some variables may have a low rank because they are redundant and yet be highly relevant [9].

3. Results and discussions

3.1. Comparison between uFilter and U-test methods

The statistical comparison between uFilter and U-test methods considered only features subsets formed by the top 10 features (empirical threshold) of each ranking. We used a total of 48 ranked subsets of image-based features for feeding four machine learning classifiers. With this, a head-to-head statistical comparison based on the mean of AUC performances over 100 runs produced inspiring results. Fig. 3 shows a boxplot graph representing the statistical comparison ($p < 0.05$) based on the mean of AUC scores between both methods for all classification schemes.

3.1.1. Results on balanced datasets

The best classification scheme for BCDR1 dataset was formed by the uFilter method and the SVM classifier (see Fig. 3b). The AUC value of 0.8369 was statistically superior to the best AUC value (0.7995) obtained by the U-test method when combined with the SVM classifier. Also, this combination was statistically better than the remaining classification schemes in the BCDR1 dataset (see Table 1).

For DDSM1 dataset, the best classification scheme was formed by the uFilter method and the SVM classifier, reaching an AUC score of 0.80. However, this result did not provide statistical evidence to be better than the combination of the U-test method and the SVM classifier, which reached an AUC score of 0.7838 (see Fig. 3b). This combination statistically outperformed only three classification schemes for DDSM1 dataset (see Table 1).

3.1.2. Results on unbalanced datasets

The best classification scheme for BCDR2 dataset was formed by the uFilter method and the FFBP neural network, reaching an

Table 1
Summary of the Wilcoxon Statistical test among all classification schemes for BCDR1 and DDSM1 datasets.

Dataset	Best scheme	AUC	Other scheme	AUC	Wilcoxon ($\alpha = 0.05$)
BCDR1	uFilter + SVM	0.8369	U-test + SVM	0.7995	$p < 0.01$
			uFilter + FFBP	0.8088	$p < 0.01$
			U-test + FFBP	0.7938	$p < 0.01$
			uFilter + LDA	0.7906	$p < 0.01$
			U-test + LDA	0.7968	$p < 0.01$
			uFilter + NB	0.7814	$p < 0.01$
			U-test + NB	0.7840	$p < 0.01$
DDSM1	uFilter + SVM	0.80	U-test + SVM	0.7838	$p = 0.1390$
			uFilter + FFBP	0.7868	$p = 0.0986$
			U-test + FFBP	0.7925	$p = 0.5149$
			uFilter + LDA	0.7567	$p < 0.01$
			U-test + LDA	0.7832	$p = 0.0624$
			uFilter + NB	0.7277	$p < 0.01$
			U-test + NB	0.7288	$p < 0.01$

Please cite this article in press as: Pérez NP, et al. Improving the Mann–Whitney statistical test for feature selection: An approach in breast cancer diagnosis on mammography. Artif Intell Med (2014), http://dx.doi.org/10.1016/j.artmed.2014.12.004

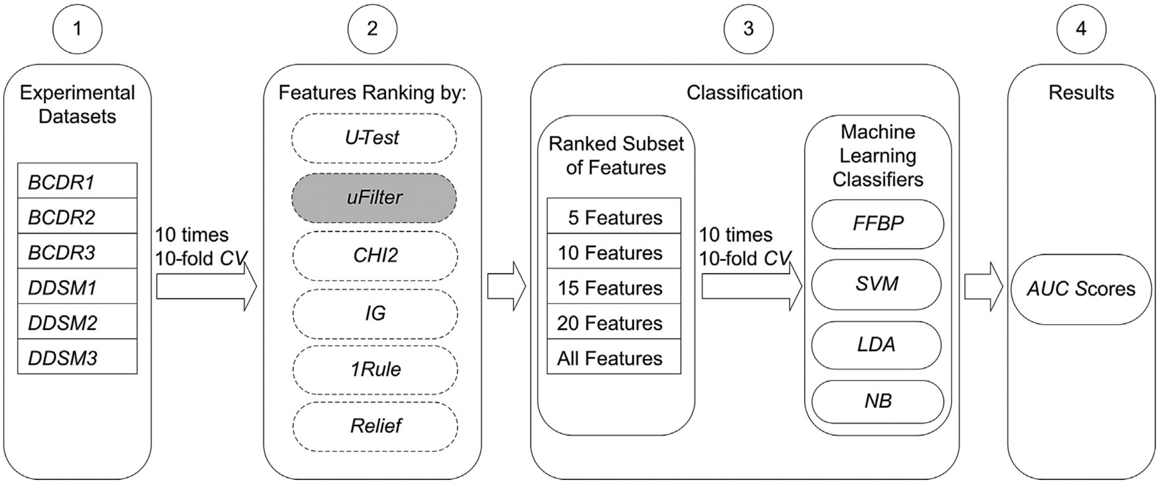

Fig. 2. Applied experimental workflow; CV means cross-validation.

AUC score of 0.8350. This result was statistically superior to the obtained result by the combination of the *U*-test method with the FFBP neural network, which provided an AUC score of 0.7578 (see Fig. 3e). In this dataset other classification schemes using the uFilter method stretched satisfactory results with no statistical difference respect to the best scheme (see Table 2).

Besides, for DDSM2 dataset the best classification performance was obtained by the combination of the uFilter and the FFBP neural network classifier; reaching AUC value of 0.8382 (see Fig. 3e). However, this result did not statistically outperform the obtained result by the combination of the *U*-test method and the FFBP neural network (AUC value of 0.8308). The comparison among all

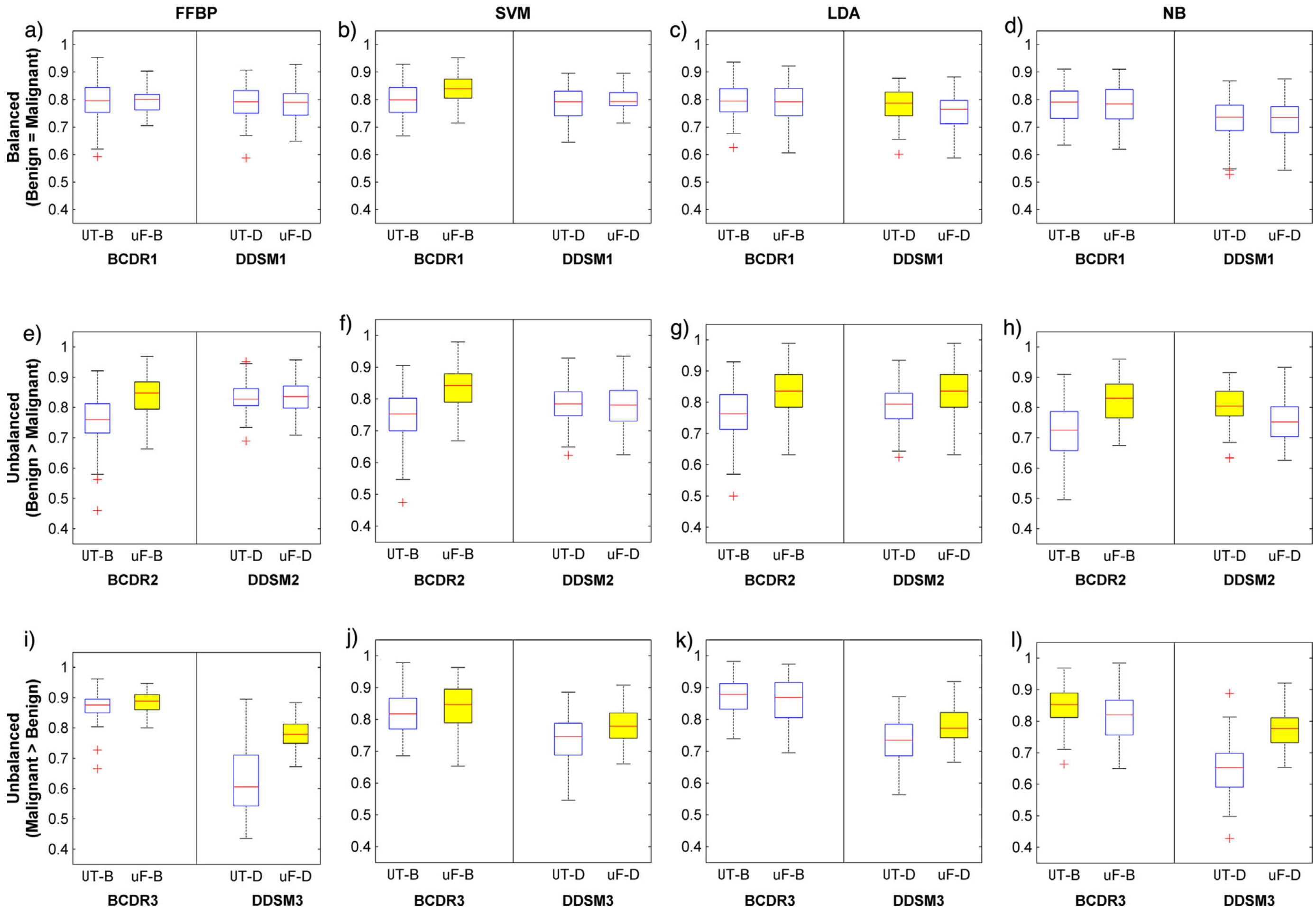

Fig. 3. Head-to-head comparison between uFilter (uF) and *U*-test (UT) methods using the top 10 features of each ranking. Filled box represents significant difference ($p < 0.05$) in the AUC performance.

Table 2
Summary of the Wilcoxon Statistical test among all classification schemes for BCDR2 and DDSM2 datasets.

Dataset	Best scheme	AUC	Other scheme	AUC	Wilcoxon (α = 0.05)
BCDR2	uFilter + FFBP	0.8350	*U*-test + FFBP	0.7578	$p < 0.01$
			uFilter + SVM	0.8332	$p = 0.6086$
			U-test + SVM	0.7482	$p < 0.01$
			uFilter + LDA	0.8296	$p = 0.5849$
			U-test + LDA	0.7613	$p < 0.01$
			uFilter + NB	0.8219	$p = 0.1546$
			U-test + NB	0.7246	$p < 0.01$
DDSM2	uFilter + FFBP	0.8382	*U*-test + FFBP	0.8308	$p = 0.4923$
			uFilter + SVM	0.7782	$p < 0.01$
			U-test + SVM	0.7844	$p < 0.01$
			uFilter + LDA	0.8296	$p = 0.7031$
			U-test + LDA	0.7881	$p < 0.01$
			uFilter + NB	0.7511	$p < 0.01$
			U-test + NB	0.8057	$p < 0.01$

classification schemes for DDSM2 dataset indicated that the best combination (higher AUC value) was almost statistically superior in all cases (see Table 2).

The best classification performance for BCDR3 dataset was provided by the combination of the uFilter method and the FFBP neural network classifier that attained, an AUC score of 0.8850 (see Fig. 3i). This result was statistically superior to the obtained result by the combination of the *U*-test method and the FFBP neural network, which achieved an AUC score of 0.87. With the exception of the scheme formed by the *U*-test method and the LDA classifier, which attained a similar AUC performance (0.8725), the best scheme statistically outperformed the remaining classification schemes (see Table 3).

In the DDSM3 dataset, the best classification scheme was formed by the combination of the uFilter method and the LDA classifier for an AUC value of 0.7819. This classification result showed significant difference with respect to the classification result provided by the combination of the *U*-test method with the LDA classifier, which reached an AUC value of 0.7328 (see Fig. 3k). Also, the best combination statistically outperformed other obtained results using the *U*-test method in the classification scheme, and does not indicated statistical evidences of being better than other uFilter combinations (see Table 3).

A head-to-head comparison between the proposed uFilter method and the classical Mann Withney *U*-test [27] (*U*-test) is well demonstrated in the experiments reported here. As it is shown in Tables 1–3, the uFilter method statistically outperformed the *U*-test method in a 50%; tied in a 37.5% and lost in a 12.5% of the 24 considered scenarios (see Fig. 3).

Table 3
Summary of the Wilcoxon Statistical test among all classification schemes for BCDR3 and DDSM3 datasets.

Dataset	Best scheme	AUC	Other scheme	AUC	Wilcoxon (α = 0.05)
BCDR3	uFilter + FFBP	0.8850	*U*-test + FFBP	0.87	$p < 0.01$
			uFilter + SVM	0.8386	$p < 0.01$
			U-test + SVM	0.8207	$p < 0.01$
			uFilter + LDA	0.8621	$p < 0.01$
			U-test + LDA	0.8725	$p = 0.2131$
			uFilter + NB	0.8152	$p < 0.01$
			U-test + NB	0.8477	$p < 0.01$
DDSM3	uFilter + LDA	0.7819	*U*-test + LDA	0.7328	$p < 0.01$
			uFilter + FFBP	0.7806	$p = 0.9386$
			U-test + FFBP	0.6266	$p < 0.01$
			uFilter + SVM	0.7795	$p = 0.8441$
			U-test + SVM	0.7393	$p < 0.01$
			uFilter + NB	0.7706	$p = 2047$
			U-test + NB	0.6467	$p < 0.01$

This circumstance could be related with the assigned weights to each feature in the ranking, e.g. in the BCDR1 dataset, the uFilter method considered the f_4 feature (perimeter) as the most relevant feature, meanwhile the *U*-test method considered it as irrelevant. A similar situation occurs with the f_{21} feature (area), which was ranked in the top five features by the uFilter and irrelevant by the *U*-test method respectively. According to the American College of Radiology [32], microcalcification lesions are tiny bright dots in the breast, and masses are very often obscure and greater than microcalcifications. Hence the perimeter and area are important features for discriminating between both lesions. It is clear that the *U*-test method underestimated both features on unbalanced datasets.

In addition, for unbalanced datasets this fact could be associated to the Mann–Whitney test criterion [27], which is based on the magnitude of the relationship between both samples (benign and malignant instances). In the BCDR2, DDSM2, BCDR3 and DDSM3 datasets most of the benign instances are greater than most of the malignant instances or vice versa and this would be evidence against random mixing. Therefore, the *U*-test method would tend to discredit the null hypothesis of identical distribution [63] and underrate the features weight (like in the balanced datasets). The opposite occur with the uFilter method, which computes the separability between both samples, independently of the number of benign and malignant instances.

3.2. *Performance of uFilter versus classical feature selection methods*

This section aims to compare the new developed uFilter method against four well known (established) feature selection methods. A total of 720 ranked subsets of image-based features were analyzed and the straightforward statistical comparison based on the mean of AUC performances over 100 runs highlighted interesting results for balanced and unbalanced datasets (see Fig. 4).

3.2.1. *Results on balanced datasets*

On the BCDR1 dataset, the best classification scheme was obtained when we combine the uFilter method and the SVM classifier using 10 features, obtaining an AUC score of 0.8369. The statistical comparison against the other feature selection methods did not provide significant difference in term of AUC scores: CHI2 discretization (AUC = 0.8325, $p = 0.6717$), IG (AUC = 0.8324, $p = 0.6725$), 1Rule (AUC = 0.8310, $p = 0.6053$) and Relief (AUC = 0.8316, $p = 0.6190$). However, the uFilter method reached this result using the top 10 features, while the other methods required a total of 20 features (see Fig. 4a).

On the DDSM1 dataset, the combination of the uFilter method and the SVM classifier using the top 10 features provided the best classification performance obtaining an AUC value of 0.8004. This result was not statistically superior to the obtained result by the CHI2 discretization (AUC = 0.7893, $p = 0.2684$), IG (AUC = 0.7893, $p = 0.2840$) and 1Rule methods (AUC = 0.79, $p = 0.3450$), but it was better than the Relief method (AUC = 0.7821, $p < 0.05$). Similar to the BCDR1 dataset, the uFilter method reached this result using the top 10 features, while the other methods required a total of 20 features.

3.2.2. *Results on unbalanced datasets*

The higher classification performance in the BCDR2 dataset was achieved by the combination of the uFilter method and the FFBP neural network classifier with a total of 10 features, obtaining an AUC value of 0.8350. However, this result was not statistically superior to the obtained results by the remaining feature selection methods using the same number of features (see Fig. 4b): CHI2 discretization (AUC = 0.8342, $p = 0.7590$), IG

Please cite this article in press as: Pérez NP, et al. Improving the Mann–Whitney statistical test for feature selection: An approach in breast cancer diagnosis on mammography. Artif Intell Med (2014), http://dx.doi.org/10.1016/j.artmed.2014.12.004

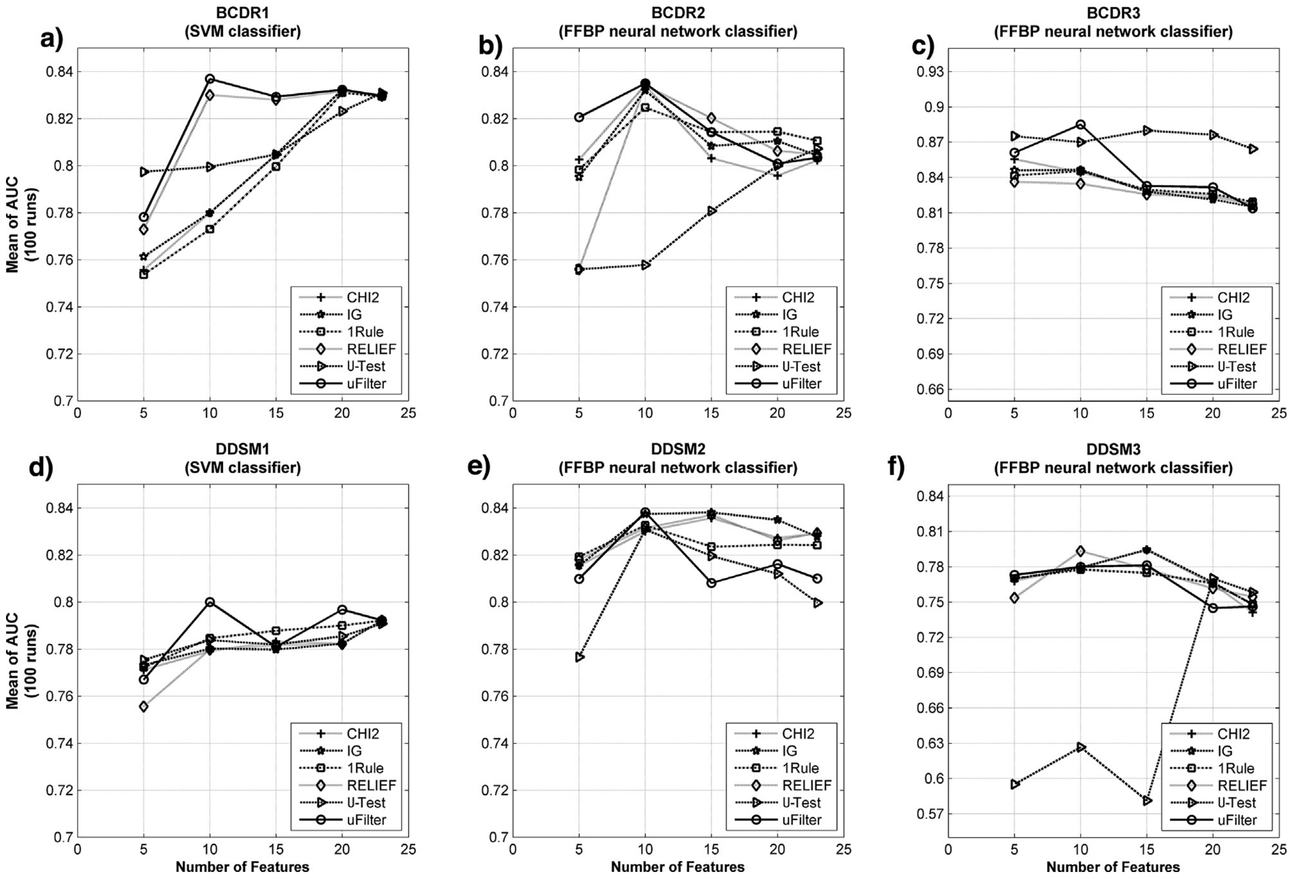

Fig. 4. Behavior of the best classification schemes when increasing the number of features on each dataset.

(AUC = 0.8320, p = 0.8022), 1Rule (AUC = 0.8259, p = 0.8259) and Relief (AUC = 0.8344, p = 0.8402). Similar to the BCDR2 dataset, the higher classification performance in the DDSM2 dataset was obtained by the combination of the uFilter method and the FFBP neural network classifier using a total of 10 features, accomplishment an AUC of 0.8382 (see Fig. 4e). This result did not provide statistical evidences of an AUC improvement respect to the CHI2 discretization (AUC = 0.8301, p = 8079), IG (AUC = 0.8374, p = 0.9076), 1Rule (AUC = 0.8326, p = 0.7470) and Relief methods (AUC = 0.8315, p = 0.3884).

On the other hand, the best classification scheme for the BCDR3 dataset was formed by the combination of the uFilter method and the FFBP neural network using a total of 10 features (see Fig. 4c). The AUC value of 0.8850 was statistically superior respect to the obtained result by the CHI2 discretization (AUC = 0.8444, p < 0.01), IG (AUC = 0.8465, p < 0.01), 1Rule (AUC = 0.8454, p < 0.01) and Relief methods (AUC = 0.8347, p < 0.01).

On the DDSM3 dataset, the best classification performance was obtained by the combination of the IG method and the FFBP neural network classifier using 15 features (AUC value of 0.7945). The AUC-based comparison against the other classification schemes using less number of features (10) indicated no significant difference in the classification performance (see Fig. 4f): CHI2 discretization (AUC = 0.7798, p = 0.0729), 1Rule (AUC = 0.7776, p = 0.0551), Relief (AUC = 0.7933, p = 0.9717) and uFilter (AUC = 0.7806, p = 0.0796).

The global comparison demonstrated that uFilter method statistically outperformed the CHI2 discretization, IG, 1Rule and Relief methods on BCDR1, DDSM1 and BCDR3 datasets, and it was statistically similar on BCDR2, DDSM2 and DDSM3 datasets while requiring less number of features. This circumstance could be related to the particular nature of employed feature selection methods and datasets respectively. We used datasets without any type of data normalization, and some methods could lead to non-reliable results e.g. the CHI2 discretization (which used the chi-square statistical test as the main evaluation function), IG (which is an entropy-based feature evaluation), and 1Rule (which is not likely to enhance the performance of classification schemes that require a search space of greater complexity) methods provided the worst results on BCDR1, DDSM1 and BCDR3 datasets (see Fig. 4). In contrast, the Relief method computes the feature's weight based on a different semantic independent of data normalization (distance to nearest hit and nearest miss), and this explains the good performance of the Relief method in almost all datasets (see Fig. 4). The satisfactory results obtained by the uFilter were expected since it is a non-parametric method and thus is tolerant to non-normalized data.

Concerning classifiers performance, results show that the selection of the most appropriate classifier is dependent on the dataset and the feature selection method (see Fig. 4). For balanced datasets, the best results were obtained with de SVM classifier; meanwhile for unbalanced datasets were obtained with the FFBP neural network classifier (see Fig. 4). These results were expected since the SVM classifier is based on the definition of an optimal hyperplane [60], and for a less complex features space (e.g. balanced datasets), it could easily find the corresponding linear decision boundary. On the other hand, for a more complex features space such as those presented on unbalanced datasets, the FFBP neural network demonstrated better capabilities of generalizing [60].

3.3. Feature relevance analysis

The results showed in previous section clearly provide experimental evidence that the uFilter method provided ranked subsets of features with higher discriminant potential. It approximates the

Please cite this article in press as: Pérez NP, et al. Improving the Mann–Whitney statistical test for feature selection: An approach in breast cancer diagnosis on mammography. Artif Intell Med (2014), http://dx.doi.org/10.1016/j.artmed.2014.12.004

set of relevant features by selecting a subset from the top 10 features of each ranking list. According to its linear time complexity in terms of the dimensionality N (total of features), the uFilter method is efficient for high-dimensional data. However, it is incapable of removing redundant features because it is an individual evaluator of features (i.e. it assigns weights according to their degrees of relevance [9]) and as long as features are considered relevant to the class, they will all be selected even though many of them are highly correlated to each other (redundant). In this case, a validation of features' subsets through a redundancy analysis is convenient.

3.3.1. Features subset validation

To efficiently find an optimal subset of features we introduced an analysis of redundancy to decrease the size of the subset of features and keeping prediction accuracy. We achieved this goal using a two-step procedure involving: (1) selecting the best subset of features for each dataset, and (2) performing the redundancy analysis based on the Pearson correlation [67] to determine and eliminate redundant features from relevant ones, and thus to produce a final optimal subset of features.

In order to correctly interpret the results, John and Kohavi in [68] defined two degrees of relevance: strong and weak. Strong relevance implies that the feature is indispensable in the sense that it cannot be removed without loss prediction accuracy. Weak relevance (redundant and non-redundant) implies that the feature can sometimes contribute to prediction accuracy. Thus, features are relevant if they are both strongly or weakly relevant and irrelevant otherwise. Irrelevant features can never contribute to prediction accuracy, by definition.

As it is shown in Fig. 4 and described in previous section, the relevance analysis based on the proposed uFilter method provided discriminant subsets of features by removing irrelevant ones. Hence, these subsets of features were used as the starting point for the redundancy analysis. Table 4 summarizes the redundancy analysis based on the Pearson correlation [67] for each selected subset of features. It should be pointed out that only higher correlation values were considered in this analysis (more than 0.5 on both positive and negative direction).

Correlated features are considered redundant (see Table 4). Therefore, only one of them in the correlated pair is selected together with the non-correlated features to form the weakly relevant subset of features. Hence each weakly relevant subset of features was used for selecting the strongly relevant ones.

In the BCDR1 dataset, the entropy (f_{12}) feature was selected as the strongly relevant feature because its absence in the final subset of features significantly decreased the AUC performance from 0.8315 to 0.791 ($p<0.01$). In the BCDR2 and DDSM3 datasets, the perimeter (f_4) feature constituted the strongly relevant feature. Its participation in the final subset significantly increased the AUC performance from 0.81 to 0.835 ($p<0.01$) and 0.7521 to 0.7806 ($p<0.01$) respectively.

For the BCDR3 dataset, the mean (f_{19}) feature was considered as the strongly relevant feature because its absence significantly reduced the classification performance from an AUC value of 0.885 to 0.8395 ($p<0.01$). Besides, in the DDSM2 dataset, the strongly relevant feature was the roughness (f_7); this feature contributed to a significantly increment of 0.07 ($p<0.01$) in the AUC performance when it is included in the final subset (AUC value of 0.8382 versus 0.7727 when is left out). Only in the DDSM1 dataset no feature appears as strongly relevant, which means that all features in this subset contributed with similar effort in the classification model. It should be pointed out that removed features can be inferred from Table 4.

According to the relevant definition of John and Kohavi [68], we put together both weakly and strongly relevant features to form the

Table 4
Summary of the redundancy analysis.

Dataset	Best subset of features	Redundant features	c-Pearson	p-Value ($\alpha=0.05$)	Weakly relevant	Strongly relevant
BCDR1	$f_4, f_{12}, f_{15}, f_{21}, f_7, f_{10}, f_3, f_6, f_{18}, f_8$	$f_{21}=f_4$ $f_{10}=f_7,f_3$ $f_3=f_7$ $f_{18}=f_6$	0.79 0.96, −0.92 −0.84 −0.62	$p<0.01$	$f_4, f_{15}^{(+)}, f_7, f_6, f_8$	f_{12}
BCDR2	$f_{14}, f_{22}, f_{21}, f_4, f_{12}, f_{15}, f_6, f_{13}, f_{11}, f_8$	$f_{14}=f_{22},f_{13}, f_{11}$ $f_{21}=f_4$ $f_{13}=f_{22}$ $f_8=f_6$	0.99, 0.56, 0.56 0.89 0.55 0.75	$p<0.01$	$f_{22}, f_{12}^{(+)}, f_{15}^{(+)}, f_6, f_{11}$	f_4
BCDR3	$f_7, f_{10}, f_3, f_4, f_{12}, f_{18}, f_{15}, f_{22}, f_{19}, f_{13}$	$f_{10}=f_7,f_3, f_{22}$ $f_3=f_7,f_{22}$ $f_{18}=f_{12}$ $f_{13}=f_7,f_{10},f_3,f_{22}$	0.97, −0.94,0.56 −0.85, −0.62 −0.75 0.50, 0.57, −0.62, 0.99	$p<0.01$	$f_7, f_4^{(+)}, f_{12}, f_{15}^{(+)}, f_{22}$	f_{19}
DDSM1	$f_9, f_{16}, f_{19}, f_{23}, f_4, f_{12}, f_{21}, f_6, f_{10}, f_{15}$	$f_{23}=f_9,f_{16},f_{19}$ $f_{21}=f_4$ $f_6=f_4, f_{15}$ $f_{15}=f_{12}$ $f_{16}=f_{19}$	0.85, 0.94, 0.94 0.93 0.56, −0.71 −0.79 0.99	$p<0.01$	$f_9^{(+)}, f_4, f_{12}, f_{10}^{(+)}, f_{19}$	–
DDSM2	$f_7, f_{19}, f_{16}, f_{23}, f_9, f_3, f_1, f_5, f_{12}, f_8$	$f_{23}=f_{19},f_{16},f_9,f_{12},f_8$ $f_9=f_{19},f_{16},f_8$ $f_{12}=f_9$	0.97, 0.98, 0.89, 0.71, 0.51 0.92, 0.92, 0.61 0.68	$p<0.01$	$f_{19}, f_{16}, f_3^{(+)}, f_1^{(+)}, f_5^{(+)}, f_8$	f_7
DDSM3	$f_9, f_4, f_{21}, f_{23}, f_{16}, f_{10}, f_{19}, f_{12}, f_{18}, f_6$	$f_{21}=f_9$ $f_{23}=f_9,f_{16},f_{19}$ $f_{16}=f_9,f_{19}$ $f_{12}=f_9,f_4,f_{18},f_6$	0.84 0.78, 0.92, 0.91 0.85, 0.99 0.60, 0.56, 0.57, 0.76	$p<0.01$	$f_9, f_{10}^{(+)}, f_{19}, f_{18}, f_6$	f_4

(+)Weakly relevant features but non-redundant; c-Pearson is the value of correlation of Pearson; p-value means whether the correlation value is significantly different from zero (i.e. are correlated).

Table 5
AUC-based statistical comparisons between the best and optimal subset of features.

Dataset	Best subset of features	AUC	Weakly + strongly	AUC	Wilcoxon ($\alpha = 0.05$)
BCDR1	$f_4, f_{12}, f_{15}, f_{21}, f_7, f_{10}, f_3, f_6, f_{18}, f_8$	0.839	$f_4, f_{15}^{(+)}, f_7, f_6, f_8, f_{12}$	0.8315	$p = 0.811$
BCDR2	$f_{14}, f_{22}, f_{21}, f_4, f_{12}, f_{15}, f_6, f_{13}, f_{11}, f_8$	0.835	$f_{22}, f_{12}^{(+)}, f_{15}^{(+)}, f_6, f_{11}, f_4$	0.8413	$p = 0.841$
BCDR3	$f_7, f_{10}, f_3, f_4, f_{12}, f_{18}, f_{15}, f_{22}, f_{19}, f_{13}$	0.885	$f_7, f_4^{(+)}, f_{12}, f_{15}^{(+)}, f_{22}, f_{19}$	0.8821	$p = 0.918$
DDSM1	$f_9, f_{16}, f_{19}, f_{23}, f_4, f_{12}, f_{21}, f_6, f_{10}, f_{15}$	0.8004	$f_9^{(+)}, f_4, f_{12}, f_{10}^{(+)}, f_{19}$	0.8001	$p = 0.982$
DDSM2	$f_7, f_{19}, f_{16}, f_{23}, f_9, f_3, f_1, f_5, f_{12}, f_8$	0.8382	$f_{19}, f_{16}, f_3^{(+)}, f_1^{(+)}, f_5^{(+)}, f_8, f_7$	0.8435	$p = 0.757$
DDSM3	$f_9, f_4, f_{21}, f_{23}, f_{16}, f_{10}, f_{19}, f_{12}, f_{18}, f_6$	0.7806	$f_9, f_{10}^{(+)}, f_{19}, f_{18}, f_6, f_4$	0.7759	$p = 0.685$

(+)Weakly relevant features but non-redundant.

optimal subset of features. These subsets were evaluated using the same machine learning classifier employed in the evaluation of its precedent subsets of features (for further comparison). Therefore, optimal subset of features for BCDR1 and DDSM1 datasets used the SVM classifier, and for BCDR2, BCDR3, DDSM2 and DDSM3 datasets used the FFBP neural network respectively. Table 5 summarizes the AUC-based statistical comparison (using the Wilcoxon statistical test [66,69]) between the best subset of features selected by the uFilter method, and its corresponding optimal subset of features after the redundancy analysis.

From Table 5, it is possible to conclude that only two optimal subsets of features provided a slight increment in terms of AUC performance, but these results were not significantly superior. Furthermore, the remaining optimal subsets of features did not provide significance difference in the AUC performance.

Concerning redundancy analysis, redundant features were detected on every dataset, which means there are some variables providing similar information to the classifier, and thus it is unnecessarily increasing the complexity of the classification model. With the exception of the DDSM1 dataset, it was possible to find the most relevant feature for all the datasets. In the case of the BCDR2 dataset, the perimeter (f_4) feature was selected as the most appropriated strongly relevant feature, however it has a unique correlation with the area (f_{21}) feature (c-Pearson value of 0.89). In this case, it is possible to interchange both features (f_4 or f_{21}) and select only one of them as the most relevant feature (see Table 4). Likewise, in the DDSM3, the perimeter (f_4) feature was selected as the most relevant feature and is correlated with the entropy (f_{12}) feature (c-Pearson value of 0.56), but the entropy (f_{12}) feature is also correlated with others features: maximum (f_9), correlation (f_{18}) and standard deviation (f_6); under this situation, the selected perimeter (f_4) feature is the only one which can be elected as the most relevant feature. This particular effect on both datasets could be explained by the c-Pearson values; the correlation value between perimeter (f_4) and area (f_{21}) was high (unique correlation) meanwhile the correlation value between perimeter (f_4) and entropy (f_{12}) was low (multi-correlation). It means that it is possible interchanging most relevant features only if there is a unique and strong correlation between them.

We considered strongly relevant features as the most important features: perimeter (f_4), entropy (f_{12}), mean (f_{19}) and roughness (f_7). They consistently appeared at least 3 times (each one) on the six optimal features subsets (see Table 5). This result was expected due to the binary classification problem (benign–malignant classes) investigated in this work. The perimeter and roughness features are considered significant shape descriptors for masses classification i.e. benign masses possess smooth, round, or oval shapes with possible macrolobulations, as opposed to malignant tumors which typically exhibit rough contours with microlobulations, spiculations, and concavities [32]. On the other hand, the entropy and mean features are more likely to be employed for microcalcification classification i.e. the entropy is a feature that represents the texture of the background tissue where the calcifications are embedded in [15,70]; meanwhile, the mean is an intensity statistics descriptor used with higher frequency [46,71] because microcalcifications are tiny brighter dots [32].

Regarding classification performances, the proposed uFilter method was able to produce subsets of features with higher discriminant potential and the redundancy analysis did not improve the prediction accuracy, but decreased the size of the subset of features without significantly decreasing the performance. This result was expected since the uFilter method is an individual evaluator of features (filter paradigm) and it ignores the feature dependencies. This is the main drawback of individual features evaluator methods as is the case of uFilter.

4. Conclusions

The new developed uFilter method performed better than the Mann–Whitney *U*-test (*U*-test) when applied to reduce and ranking features in binary classification problems. uFilter was validated using several machine learning algorithms such as FFBP neural network, SVM, LDA and NB classifiers over six different (balanced and unbalanced) datasets representative of two different breast cancer repositories. A head-to-head comparison proved that the uFilter method significantly outperformed the *U*-test method for almost all of the classification schemes. It was superior in 50%; tied in a 37.5% and lost in a 12.5% of the 24 comparative scenarios. Moreover, a global comparison against other four well known feature selection methods (CHI2 discretization, IG, 1Rule and Relief) demonstrated that uFilter statistically outperformed the remaining methods on several datasets (BCDR1, DDSM1 and BCDR3), and it was statistically similar on the BCDR2, DDSM2 and DDSM3 datasets while requiring less number of features. The uFilter method revealed competitive and appealing cost-effectiveness results on selecting relevant features, as a support tool for breast cancer CADx methods especially in unbalanced datasets contexts. Finally, the redundancy analysis as a complementary step to the uFilter method provided us an effective way for finding optimal subsets of features without decreasing the classification performances.

Future work will be aimed to: (1) increasing the number of features in benchmarking breast cancer datasets; (2) exploring the performance of uFilter in other knowledge domains and (3) extending uFilter allowing it to be used on multiclass classification problems.

Conflict of interest

The authors declare that they do not have any conflict of interest.

Acknowledgments

MSc. Pérez acknowledges “Fundação para a Ciência e a Tecnologia (grant SFRH/BD/48896/2008)” for financial support. Prof. Guevara acknowledges the Cloud Thinking project (CENTRO-07-ST24-FEDER-002031), co-funded by QREN, “Mais Centro” program.

Please cite this article in press as: Pérez NP, et al. Improving the Mann–Whitney statistical test for feature selection: An approach in breast cancer diagnosis on mammography. Artif Intell Med (2014), http://dx.doi.org/10.1016/j.artmed.2014.12.004

Algorithm 1. uFilter method

```
1. Let $F = \{f_1, f_2, \ldots, f_t\}$ a set of features with size $t$;
2. Let $f_i = \{I_{c,1}, I_{c,2}, \ldots, I_{c,n}\}$ a set of instances with size $n$, where $I_{c,j}$ is the
   value of the feature $f_i$ for the instance $j$, and $c$ denotes the class value ($B$
   or $M$);
3. For each $f_i$
   a. Initial weight of the feature $w_i = 0$;
   b. Sort ($f_i{}'$, ascendant');
   c. Perform the tie analysis of resultant in b:
      $f_i{}'$ = avg(position of tied elements);
   d. Compute the summation of benign and malignant instances based on
   Eqs. (1) and (2);
   e. Compute $u$-values based on Eqs. (3) and (4)
   f. Compute $Z$-indicators based on Eqs. (5) and (6)
   g. Updating the weight of the feature based on Eq. (7)
4. End for
5. Output ranking Sort($w$,' descendant');
```

References

[1] Devijver PA, Kittler J. Pattern recognition: a statistical approach. London, UK: Prentice/Hall International; 1982.

[2] Guyon I, Elisseeff A. An introduction to feature extraction. In: Guyon I, Nikravesh M, Gunn S, Zadeh L, editors. Feature extraction, vol. 207. Springer Berlin/Heidelberg; 2006. p. 1–25, http://dx.doi.org/10.1007/978-3-540-35488-8_1.

[3] Blanco M, Coello J, Iturriaga H, Maspoch S, de la Pezuela C. Near-infrared spectroscopy in the pharmaceutical industry. Analyst 1998;123:135R–50R, http://dx.doi.org/10.1039/a802531b.

[4] Koehn FE, Carter GT. The evolving role of natural products in drug discovery. Nat Rev Drug Discov 2005;4:206–20, http://dx.doi.org/10.1038/nrd1657.

[5] Hashemi H, Tax DMJ, Duin RPW, Javaherian A, de Groot P. Gas chimney detection based on improving the performance of combined multilayer perceptron and support vector classifier. Nonlinear Process Geophys 2008;15:863–71, http://dx.doi.org/10.5194/npg-15-863-2008.

[6] Kaifeng Y, Wenkai L, Wenlong D, Shanwen Z, Huanqin X, Yanda L. Hydrocarbon prediction method based on Svm feature selection. Nat Gas Ind 2004;24:36–8.

[7] Chanwoo K, Stern RM. Feature extraction for robust speech recognition based on maximizing the sharpness of the power distribution and on power flooring. In: 2010 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP). 2010. p. 4574–7, http://dx.doi.org/10.1109/ICASSP.2010.5495570.

[8] Pei Y, Essa I, Starner T, Rehg JM. Discriminative feature selection for hidden Markov models using Segmental Boosting. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2008). 2008. p. 2001–4, http://dx.doi.org/10.1109/ICASSP.2008.4518031.

[9] Guyon I, Elisseeff A. An introduction to variable and feature selection. J Mach Learn Res 2003;3:1157–82.

[10] Yu L, Liu H. Efficient feature selection via analysis of relevance and redundancy. J Mach Learn Res 2004;5:1205–24.

[11] Holmberg M, Gustafsson F, Hornsten EG, Winquist F, Nilsson LE, Ljung L, et al. Bacteria classification based on feature extraction from sensor data. Biotechnol Tech 1998;12:319–24, http://dx.doi.org/10.1023/A:1008862617082.

[12] Dutta R, Hines EL, Gardner JW, Boilot P. Bacteria classification using Cyranose 320 electronic nose. BioMed Eng Online 2002;1:4, http://dx.doi.org/10.1186/1475-925x-1-4.

[13] López Y, Novoa A, Guevara M, Silva A. Breast cancer diagnosis based on a suitable combination of deformable models and artificial neural networks techniques. In: Rueda L, Mery D, Kittler J, editors. Progress in pattern recognition image analysis and applications, vol. 4756/2008. Berlin/Heidelberg: Springer; 2008. p. 803–11, http://dx.doi.org/10.1007/978-3-540-76725-1_83.

[14] Ghazavi SN, Liao TW. Medical data mining by fuzzy modeling with selected features. Artif Intell Med 2008;43:195–206, http://dx.doi.org/10.1016/j.artmed.2008.04.004.

[15] Soltanian-Zadeh H, Rafiee-Rad F, Pourabdollah-Nejad SD. Comparison of multiwavelet, wavelet, haralick, and shape features for microcalcification classification in mammograms. Pattern Recognit 2004;37:1973–86, http://dx.doi.org/10.1016/j.patcog.2003.03.001.

[16] Wei J, Sahiner B, Hadjiiski LM, Chan H-P, Petrick N, Helvie MA, et al. Computer-aided detection of breast masses on full field digital mammograms. Med Phys 2005;32:2827–38, http://dx.doi.org/10.1118/1.1997327.

[17] Lee SK, Chung PC, Chang CI, Lo CS, Lee T, Hsu GC, et al. Classification of clustered microcalcifications using a Shape Cognitron neural network. Neural Netw 2003;16:121–32, http://dx.doi.org/10.1016/S0893-6080(02)00164-8.

[18] Saeys Y, Inza I, Larranaga P. A review of feature selection techniques in bioinformatics. Bioinformatics 2007;23:2507–17, http://dx.doi.org/10.1093/bioinformatics/btm344.

[19] Setiono R, Liu H. CHI2: feature selection and discretization of numeric attributes. In: IEEE Proceedings of the Seventh International Conference on Tools with Artificial Intelligence. 1995. p. 388–91, http://dx.doi.org/10.1109/TAI.1995.479783.

[20] Liu H, Li J, Wong L. A comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns. Genome Inform 2002;13:51–60.

[21] Press WH, Flannery BP, Teukolsky SA, Vetterling WT. Numerical recipes in C. Cambridge: Cambridge University Press; 1988.

[22] Jain AK, Chandrasekaran B. 39 Dimensionality and sample size considerations in pattern recognition practice. In: Krishnaiah PR, Kanal LN, editors. Handbook of statistics, vol. 2. Elsevier; 1982. p. 835–55, http://dx.doi.org/10.1016/S0169-7161(82)02042-2.

[23] Hall MA, Smith LA. Feature selection for machine learning: comparing a correlation-based filter approach to the wrapper. In: Proceedings of the Twelfth International Florida Artificial Intelligence Research Society Conference. 1999.

[24] Koller D, Sahami M. Toward optimal feature selection. Stanford InfoLab, Technical report 1996–77; 1996.

[25] Kira K, Rendell LA. A practical approach to feature selection. In: Sleeman D, Edwards P, editors. ML92 Proceedings of the ninth international workshop on Machine learning. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.; 1992. p. 249–56.

[26] Prados J, Kalousis A, Sanchez JC, Allard L, Carrette O, Hilario M. Mining mass spectra for diagnosis and biomarker discovery of cerebral accidents. Proteomics 2004;4:2320–32, http://dx.doi.org/10.1002/pmic.200400857.

[27] Kirk RE. Statistics: an introduction. fifth ed. Belmont, CA, USA: Thomson/Wadsworth; 2007.

[28] Ramos-Pollan R, Guevara-Lopez MA, Suarez-Ortega C, Diaz-Herrero G, Franco-Valiente JM, Rubio-Del-Solar M, et al. Discovering mammography-based machine learning classifiers for breast cancer diagnosis. J Med Syst 2012;36:2259–69, http://dx.doi.org/10.1007/s10916-011-9693-2.

[29] Breast cancer digital repository; 2014. http://bcdr.inegi.up.pt/ [accessed 02.01.14].

[30] de Oliveira JE, Machado AM, Chavez GC, Lopes AP, Deserno TM, Araujo Ade A. MammoSys: a content-based image retrieval system using breast density patterns. Comput Methods Progr Biomed 2010;99:289–97, http://dx.doi.org/10.1016/j.cmpb.2010.01.005.

[31] Heath M, Bowyer K, Kopans D, Moore R, Kegelmeyer WP. The digital database for screening mammography. In: Yaffe MJ, editor. Proceedings of the Fifth International Workshop on Digital Mammography. Medical Physics Publishing; 2001. p. 212–8.

[32] Committee. American College of Radiology (ACR) ACR BIRADS – mammography. In: A.C.o. Radiology, editor. ACR breast imaging reporting and data system, breast imaging atlas. Reston, VA: American College of Radiology; 2003.

[33] Haralick RM, Shanmuga K, Dinstein I. Textural features for image classification. IEEE Trans Syst Man Cybern 1973;Smc3:610–21, http://dx.doi.org/10.1109/Tsmc.1973.4309314.

[34] López Y, Novoa A, Guevara M, Quintana N, Silva A. Computer aided diagnosis system to detect breast cancer pathological lesions. In: Ruiz-Shulcloper J, Kropatsch W, editors. Progress in pattern recognition, image analysis and applications, vol. 5197. Springer Berlin/Heidelberg; 2008. p. 453–60, http://dx.doi.org/10.1007/978-3-540-85920-8_56.

[35] Dash M, Liu H. Feature selection for classification. Intell Data Anal 1997;1:131–56, http://dx.doi.org/10.3233/IDA-1997-1302.

[36] Holte RC. Very simple classification rules perform well on most commonly used datasets. Mach Learn 1993;11:63–91, http://dx.doi.org/10.1023/A:1022631118932.

[37] Malar E, Kandaswamy A, Chakravarthy D, Giri Dharan A. A novel approach for detection and classification of mammographic microcalcifications using wavelet analysis and extreme learning machine. Comput Biol Med 2012;42:898–905, http://dx.doi.org/10.1016/j.compbiomed.2012.07.001.

[38] Verma B, Panchal R. Neural networks for the classification of benign and malignant patters in digital mammograms. In: Sugumaran V, editor. Intelligent information technologies: concepts, methodologies, tools, and applications. Hershey, NY, USA: IGI Global; 2008. p. 947–67, http://dx.doi.org/10.4018/978-1-59904-941-0.ch056.

[39] Bellotti R, De Carlo F, Tangaro S, Gargano G, Maggipinto G, Castellano M, et al. A completely automated CAD system for mass detection in a large mammographic database. Med Phys 2006;33:3066–75.

[40] Papadopoulos A, Fotiadis DI, Likas A. Characterization of clustered microcalcifications in digitized mammograms using neural networks and support vector machines. Artif Intell Med 2005;34:141–50, http://dx.doi.org/10.1016/j.artmed.2004.10.001.

[41] Pérez N, Guevara MA, Silva A, Ramos I, Loureiro J. Improving the performance of machine learning classifiers for Breast Cancer diagnosis based on feature selection. In: Ganzha M, Maciaszek L, Paprzycki M, editors. IEEE 2014 Federated Conference on Computer Science and Information Systems (FedCSIS). 2014. p. 209–17, http://dx.doi.org/10.15439/2014F249.

[42] Pérez N, Guevara MA, Silva A. Improving breast cancer classification with mammography, supported on an appropriate variable selection analysis. In: Novak CL, Aylward S, editors. SPIE medical imaging 2013. Lake Buena Vista (Orlando Area), Florida, USA: International Society for Optics and Photonics; 2013. p. 867022-1–14, http://dx.doi.org/10.1117/12.2007912.

[43] Ping Z, Verma B, Kuldeep K. A neural-genetic algorithm for feature selection and breast abnormality classification in digital mammography. In: IEEE

Please cite this article in press as: Pérez NP, et al. Improving the Mann–Whitney statistical test for feature selection: An approach in breast cancer diagnosis on mammography. Artif Intell Med (2014), http://dx.doi.org/10.1016/j.artmed.2014.12.004

International Joint Conference on Neural Networks, vol. 3. 2004. p. 2303–8, http://dx.doi.org/10.1109/IJCNN.2004.1380985.

[44] Mavroforakis ME, Georgiou HV, Dimitropoulos N, Cavouras D, Theodoridis S. Mammographic masses characterization based on localized texture and dataset fractal analysis using linear, neural and support vector machine classifiers. Artif Intell Med 2006;37:145–62, http://dx.doi.org/10.1016/j.artmed.2006.03.002.

[45] Pérez N, Guevara MA, Silva A. Evaluation of features selection methods for breast cancer classification. In: Silva J, Vaz M, editors. 15th International Conference on Experimental Mechanics (ICEM15). Portugal: Porto; 2012.

[46] Fu JC, Lee SK, Wong ST, Yeh JY, Wang AH, Wu HK. Image segmentation feature selection and pattern classification for mammographic microcalcifications. Comput Med Imaging Graphics 2005;29:419–29, http://dx.doi.org/10.1016/j.compmedimag.2005.03.002.

[47] Shi J, Sahiner B, Chan HP, Ge J, Hadjiiski L, Helvie MA, et al. Characterization of mammographic masses based on level set segmentation with new image features and patient information. Med Phys 2008;35:280–90, http://dx.doi.org/10.1118/1.2820630.

[48] Jesneck JL, Lo JY, Baker JA. Breast mass lesions: computer-aided diagnosis models with mammographic and sonographic descriptors. Radiology 2007;244:390–8, http://dx.doi.org/10.1148/radiol.2442060712.

[49] Gupta S, Chyn PF, Markey MK. Breast cancer CADx based on BI-RAds descriptors from two mammographic views. Med Phys 2006;33:1810–7.

[50] Catarious Jr DM, Baydush AH, Floyd Jr CE. Incorporation of an iterative, linear segmentation routine into a mammographic mass CAD system. Med Phys 2004;31:1512–20.

[51] Moura D, Guevara López M. An evaluation of image descriptors combined with clinical data for breast cancer diagnosis. Int J Comput Assist Radiol Surg 2013;8:561–74, http://dx.doi.org/10.1007/s11548-013-0838-2.

[52] Salama GI, Abdelhalim M, Zeid MA-e. Breast cancer diagnosis on three different datasets using multi-classifiers. Breast Cancer (WDBC) 2012;32:2.

[53] Christobel A. An empirical comparison of data mining classification methods. Int J Comput Inf Syst 2011;3(2):24–8.

[54] Kim M, Ryu J. Optimized fuzzy classification using genetic algorithm. In: Wang L, Jin Y, editors. Fuzzy systems and knowledge discovery, vol. 3613. Springer Berlin Heidelberg; 2005. p. 392–401, http://dx.doi.org/10.1007/11539506_51.

[55] Abonyi J, Szeifert F. Supervised fuzzy clustering for the identification of fuzzy classifiers. Pattern Recognit Lett 2003;24:2195–207, http://dx.doi.org/10.1016/S0167-8655(03)00047-3.

[56] Xu W, Xia S, Xie H. Application of CMAC-based networks on medical image classification. In: Yin F-L, Wang J, Guo C, editors. Advances in neural networks – ISNN 2004, vol. 3173. Springer Berlin Heidelberg; 2004. p. 953–8, http://dx.doi.org/10.1007/978-3-540-28647-9_157.

[57] Song H, Lee S, Kim D, Park G. New methodology of computer aided diagnostic system on breast cancer. In: Wang J, Liao X-F, Yi Z, editors. Advances in neural networks – ISNN 2005, vol. 3498. Springer Berlin Heidelberg; 2005. p. 780–9, http://dx.doi.org/10.1007/11427469_124.

[58] Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA data mining software: an update. ACM SIGKDD Explor Newslett 2009;11:10–8, http://dx.doi.org/10.1145/1656274.1656278.

[59] Garcıía López F, Garcıía Torres M, Melián Batista B, Moreno Pérez JA, Moreno-Vega JM. Solving feature subset selection problem by a Parallel Scatter Search. Eur J Oper Res 2006;169:477–89, http://dx.doi.org/10.1016/j.ejor.2004.08.010.

[60] Hwang J-N. Introduction to neural networks for signal processing. In: Hu YH, Hwang J-N, editors. Handbook of neural network signal processing. Boca Raton, Florida, USA: CRC Press; 2001. p. 408.

[61] Duda RO, Hart PE, Stork DG. Pattern classification. 2nd ed. New York, NY, USA: Wiley-Interscience; 2000.

[62] Wang S, Summers RM. Machine learning and radiology. Med Image Anal 2012;16:933–51, http://dx.doi.org/10.1016/j.media.2012.02.005.

[63] Marques de Sá JP, editor. Estimating data parameters applied statistics using SPSS, STATISTICA, MATLAB and R. Springer Berlin/Heidelberg; 2007. p. 81–109, http://dx.doi.org/10.1007/978-3-540-71972-4_3.

[64] Hastie T, Tibshirani R, Friedman J, Franklin J. The elements of statistical learning: data mining, inference and prediction. Math Intell 2005;27:83–5.

[65] Mann HB, Whitney DR. On a test of whether one of two random variables is stochastically larger than the other. Ann Math Stat 1947;18:50–60, http://dx.doi.org/10.1214/aoms/1177730491.

[66] Demsar J. Statistical comparisons of classifiers over multiple data sets. J Mach Learn Res 2006;7:1–30.

[67] Gibbons J, Chakraborti S. Nonparametric statistical inference. In: Lovric M, editor. International encyclopedia of statistical science. Springer Berlin Heidelberg; 2011. p. 977–9, http://dx.doi.org/10.1007/978-3-642-04898-2_420.

[68] John GH, Kohavi R, Pfleger K. Irrelevant features and the subset selection problem. In: Cohen WW, Hirsh H, editors. Machine Learning, Proceedings of the Eleventh International Conference. New Brunswick, NJ, USA: Morgan Kaufmann, Rutgers University; 1994. p. 121–9.

[69] Hollander M, Wolfe D. Nonparametric statistical methods. 3rd ed. Hoboken, New Jersey, USA: John Wiley & Sons; 2013.

[70] AbuBaker A, Qahwaji R, Ipson S. Texture-based feature extraction for the microcalcification from digital mammogram images. In: IEEE International Conference on Signal Processing and Communications (ICSPC 2007). 2007. p. 896–9, http://dx.doi.org/10.1109/ICSPC.2007.4728464.

[71] Mohanty A, Senapati M, Lenka S. An improved data mining technique for classification and detection of breast cancer from mammograms. Neural Comput Appl 2013;22:303–10, http://dx.doi.org/10.1007/s00521-012-0834-4.