Expert Systems with Applications 38 (2011) 4600–4607

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Feature selection using fuzzy entropy measures with similarity classifier

Pasi Luukka [*]

Laboratory of Applied Mathematics, Lappeenranta University of Technology, P.O. Box 20, FIN-53851 Lappeenranta, Finland

ARTICLE INFO

Keywords:
Feature selection
Fuzzy measures
Similarity
Classification
Medical data

ABSTRACT

Feature selection plays an important role in classification for several reasons. First it can simplify the model and this way computational cost can be reduced and also when the model is taken for practical use fewer inputs are needed which means in practice that fewer measurements from new samples are needed. Second by removing insignificant features from the data set one can also make the model more transparent and more comprehensible, providing better explanation of suggested diagnosis, which is an important requirement in medical applications. Feature selection process can also reduce noise and this way enhance the classification accuracy. In this article, feature selection method based on fuzzy entropy measures is introduced and it is tested together with similarity classifier. Model was tested with four medical data sets which were, dermatology, Pima-Indian diabetes, breast cancer and Parkinsons data set. With all the four data sets, we managed to get quite good results by using fewer features that in the original data sets. Also with Parkinsons and dermatology data sets, classification accuracy was managed to enhance significantly this way. Mean classification accuracy with Parkinsons data set being 85.03% with only two features from original 22. With dermatology data set, mean accuracy of 98.28% was achieved using 29 features instead of 34 original features. Results can be considered quite good.

© 2010 Elsevier Ltd. All rights reserved.

1. Introduction

Machine learning-based classification techniques provide support for the decision-making process in many areas of health care, including prognosis, diagnosis, screening, etc. Accuracy is very important in classifiers, especially in medical applications. A high percentage of false negatives in screening systems increases the risk of patients not getting the attention they need. On the other hand, a high false alarm rate causes unwanted worries and increases the load on medical resources. In quest for higher classification accuracies, feature subset selection has been used for data reduction in areas characterized by high dimensionality due to the large numbers of available features, e.g. in seismic data processing (Hoffman, Hoogenboezem, Van der Merwe, & Tollig, 1998), remote sensing (Yu, De Backer, & Scheunders, 2000), drug design (Ozdemir et al., 2001), speech recognition (Abdulla & Kasabov, 2003) and image segmentation (Matsui & Kosugi, 1999). Feature selection is expected to improve classification performance, particularly in situations characterized by the high data dimensionality problem caused by relatively few training examples compared to a large number of measured features. This type of situation arises frequently in medicine where considerations of risk, time, difficulty, cost and inconvenience may limit the number of training examples, while the number of disease markers increases rapidly over the years (Lewin, 2000).

Even if no significant improvements in classification accuracy are achieved, reducing number of features still has many advantages. These are e.g. reducing the number of measurements required, shortening training and execution times, and improving model compactness, transparency, and interpretability. Reducing the number of features to be measured for model implementation makes screening tests faster, more convenient and less costly. Simpler models with fewer inputs are also more transparent and more comprehensible, providing better explanation of suggested diagnosis, which is an important requirement in medical applications. Fewer model inputs result in simpler models that train and execute faster and allow training on smaller data sets without the risk of overfitting. Discarding irrelevant and redundant features reduces noise and spurious correlations with the output and avoids the problems of collinearity between inputs. Feature reduction has been applied to several areas in medicine. These are e.g. classification of ultrasound liver tissues using the wavelet transform (Lee, Chen, & Hsieh, 2003), classification of EEG signals for operating brain–computer interfaces (Garret, Peterson, Anderson, & Thaut, 2003), detection of mass lesions in digital mammograms (Kupinski & Giger, 1997), classification of hepatic lesions from computed tomography images (Gletsos et al., 2003), segmenting digital chest radiographs (McNitt-Gray, Huang, & Sayre, 1995), etc.

Techniques for feature subset selection can be classified into following categories: embedded, filter and wrapper techniques

[*] Tel.: +358 503694108.
E-mail address: pasi.luukka@lut.fi

0957-4174/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2010.09.133

(Blum & Langley, 1997). In embedded techniques, feature selection can be considered to be a part of the learning itself. By testing the values of certain features, algorithms seek to split the training data into subsets. Filter and wrapper techniques on the other hand perform feature selection as a preprocessing step prior to the classifier. There the objective being to select an optimum feature subset that serves as an input to the learning algorithm. Filter techniques do not use the learning mechanism for the feature selection. They are designed to filter out undesirable features through checking data consistency and eliminating features whose information content is represented by others. The filter approach evaluates and selects feature subsets based on general characteristics of data, and usually also some statistical analysis without employing any learning model. One advantage of filter techniques is that since they do not use the learning algorithm, they are usually fast and therefore suitable for use with large data sets. Also they are easily applicable to various learning techniques. Wrapper techniques (Kohavi & John, 1997) search for an optimal feature subset through testing the performance of candidate subsets using the learning algorithm. The wrapper technique involves a learning model and uses its performance as the evaluation criterion. The wrapper approach is known to be more accurate compared to the filter approach and it is computationally more expensive. As the learning algorithm is called repeatedly, wrapper methods are slower than filter methods and do not scale up well to large, high-dimensional data sets. To try to overcome this limitation, usually a fast and simple learning algorithm is used with them e.g. nearest-neighbor classifier. Wrapper feature selections are unique to the learning algorithm used, and the process should be repeated for a different learning algorithm. Besides these techniques also hybrid approaches are developed. These are usually combination of filter and wrapper technique and are designed to trade accuracy with computational speed by applying a wrapper technique to only those subsets pre-selected by a filter technique. Strategies used for searching the feature space include sequential feature selection (SFS) methods (Aha & Bankert, 1996), which are either forward sequential search (FSS) or backward sequential search (BSS). FSS starts with an empty set, adding single features that best improve performance criteria. BSS starts with the full feature set and sequentially removes features that best improve performance criteria.

In this article as a classifier, we are using similarity-based classification procedure (Luukka & Leppälampi, 2006; Luukka, Saastamoinen, & Könönen, 2001), and for the feature selection process, we propose a new feature selection technique based on fuzzy entropy measures (De Luca & Termini, 1971; Parkash, Sharma, & Mahajan, 2008). Data sets used in this experiment were taken from a UCI-Repository of Machine Learning Database (Newman, Hettich, Blake, & Merz, 2007). Chosen data sets were dermatology, Pima-Indian diabetes, breast cancer and Parkinsons data set. Classifier and feature selection method was implemented with *MATLAB*™-software.

2. Data sets

Next, a short description of the data sets used in this article is given. Data sets were taken from UCI machine learning data repository (Newman et al., 2007) where they are freely available. The fundamental properties of the data sets are shown in Table 1.

2.1. Breast cancer data set

This data set was created by Dr. Wolberg, and the purpose was to accurately diagnose breast masses based solely on Fine Needle Aspiration (FNA). He identified nine visually assessed characteristics of an FNA sample which he considered relevant to the diagnosis. The resulting data set is well known as the Wisconsin Breast Cancer Data. The nine variables used to predict benign or malignant cases were as follows: (1) Clump Thickness, (2) Uniformity of Cell Size, (3) Uniformity of Cell Shape, (4) Marginal Adhesion, (5) Single Epithelial Cell Size, (6) Bare Nuclei, (7) Bland Chromatin, (8) Normal Nucleoli and (9) Mitoses.

Table 1
Test data sets and their properties.

Data set	Nb. classes	Nb. features	Nb. cases
Dermatology	6	34	366
Pima-Indians	2	8	768
Breast cancer	2	9	699
Parkinsons	2	22	197

2.2. PIMA-Indians

The Pima-Indian data set concerns the presence or absence of diabetes among Pima-Indian women living near Phoenix, Arizona. There are eight covariates: number of pregnancies; plasma glucose concentration; diastolic blood pressure (mmHg); tricep skin fold thickness (mm); serum insulin (μU/ml); body mass index (kg m^{-2}); diabetes pedigree function; and age in years.

2.3. Dermatology data set

The data set comes from Gazi University and Bilkent University and was donated by N. Ilker and H.A. Güvenir. This data set contains 34 attributes and contains 366 instances. Attributes and distribution due to class variable of this data set are given in Table 2. The erythemato-squamous diseases are psoriasis, seboreic dermatitis, lichen planus, pityriasis rosea, cronic dermatitis and pityriasis rubra pilaris. These diseases are frequently seen in outpatient dermatology departments. They all share the clinical features of erythema and scaling with slight variations and this makes the differential diagnosis of erythemato-squamous diseases difficult. All the diseases look very much alike with erythema and scaling. When inspected more carefully, some patients have the typical clinical features of the disease at predilection sites (localizations of the skin which a disease prefers) while another group has typical localizations. Another difficulty for differential diagnosis is that a disease may show the histopathological features of another disease in the early stages and may have the characteristic features in the following stages. Furthermore, some samples show the typical histopathological features of the disease while some do not (Übeyli & Güler, 2005).

2.4. Parkinsons data set

The data set was created by Max Little of the University of Oxford. Data set is composed of a range of biomedical voice measurements from healthy people and people with Parkinson's disease (PD). Each column in the table is a particular voice measure, and each row corresponds one of 195 voice recording from people who participated in collection of this data. The main aim of the data is to discriminate healthy people from those with PD.

3. Classification procedure

3.1. Fuzzy entropy measures and their usage with similarity classifier

In many cases, it is of interest to have suitable measures of impreciseness and vagueness, so called fuzziness measures which

Table 2
Class distribution of dermatology data set.

Class	Attributes clinical	Histopathological
Psoriasis (111)	Att. 1: erythema	Att. 12: melanin incontinence
Seboreic dermatitis (60)	Att.2: scaling	Att. 13: eosinophils in infiltrate
Lichen planus (71)	Att. 3: definite borders	Att. 14: PNL infiltrate
Pityriasis rosea (48)	Att. 4: itching	Att. 15: fibrosis of the papillary dermis
Cronic dermatitis (48)	Att. 5: koebner phenomenon	Att. 16: exocytosis
Pityarisis rubra pilaris (20)	Att. 6: polygonal papules	Att. 17: acanthosis
	Att. 7: follicular papules	Att. 18: hyperkeratosis
	Att. 8: oral mucosal involvement	Att. 19: parakeratosis
	Att. 9: knee and elbow involvement	Att. 20: clubbing of the rete ridges
	Att. 10: scalp involvement	Att. 21: elongation of the rete ridges
	Att. 11: family history	Att. 22: thinning of the suprapapillary epidermis
	Att. 34: age	Att. 23: pongiform pustule
		Att. 24: munro microabscess
		Att. 25: focal hypergranulosis
		Att. 26: disappearance of the granular layer
		Att. 27: vascularization and damage of basal layer
		Att. 28: spongiosis
		Att. 29: saw-tooth appearance of retes
		Att. 30: follicular horn plug
		Att. 31: perifollicular parakeratosis
		Att. 32: inflammatory mononuclear infiltrate
		Att. 33: band-like infiltrate

give an answer to the question: How far is a given fuzzy set from well-defined crisp reference sets? (Bandemer & Näther, 1992).

The specificity of fuzzy sets (Zadeh, 1965) is to capture the idea of partial membership. Taking into consideration, the concept of fuzzy sets, De Luca and Termini (1971) suggested that corresponding to Shannon (1948) probabilistic entropy, the measure of fuzzy entropy should be

$$H_1(A) = -\sum_{j=1}^{n}(\mu_A(x_j)log\mu_A(x_j) + (1-\mu_A(x_j))log(1-\mu_A(x_j))) \qquad (1)$$

where $\mu_A(x_j)$ are the fuzzy values. This fuzzy entropy measure is considered to be a fuzziness measure (Bandemer & Näther, 1992), and it evaluate global deviations from the type of ordinary sets, i.e. any crisp set A_0 leads to $h(A_0) = 0$. Note that the fuzzy set A with $\mu_A(x) = 0.5$ plays the role of maximum element of the ordering defined by H. Newer fuzzy entropy measures were introduced by Parkash et al. Parkash et al. (2008) where fuzzy entropies were defined as

$$H_2(A;w) = \sum_{j=1}^{n} w_j\left(\sin\frac{\pi\mu_A(x_j)}{2} + \sin\frac{\pi(1-\mu_A(x_j))}{2} - 1\right) \qquad (2)$$

and

$$H_3(A;w) = \sum_{j=1}^{n} w_j\left(\cos\frac{\pi\mu_A(x_j)}{2} + \cos\frac{\pi(1-\mu_A(x_j))}{2} - 1\right) \qquad (3)$$

These fuzzy entropy measures were used in feature selection process. Next, we go more into details of that subject. The main basic principle in similarity classifier is that one first creates the ideal vectors $\mathbf{v}_i = (v_i(f_1),\ldots,v_i(f_t))$ that represents the class i as well as possible. This vector can be user defined or calculated from some sample set X_i of vectors $\mathbf{x} = (x(f_1),\ldots,x(f_t))$ which are known to belong to class C_i. Simple way to do this is e.g. to use the generalized mean. After these ideal vectors have been calculated, one calculates the similarities $S\langle\mathbf{x},\mathbf{v}\rangle$ between the sample $\mathbf{x}$ one wants to classify and the ideal vectors $\mathbf{v}$. The decision to which class the sample belongs is made according to the similarity value between the sample and ideal vector. More about this process is provided in next subsection. Now in the ideal case if the sample belongs to class i, we get the similarity value between the ideal vector and sample being $S\langle\mathbf{x},\mathbf{v}\rangle = 1$. If the sample does not belong to this class in ideal case, we get zero from the similarity value. Now when we calculate the similarities between sample and ideal vector, we can do it so that we get j similarities where j is the number of features. At this point comes the idea of using fuzzy entropy measures to calculate the relevance of the features. If we calculate the fuzzy entropy (1) values now with $\mu_A(x_j)$ being similarity values, we get low entropy values if we get high similarity values and if we get similarity values close to 0.5, we are getting high entropy values. Using this underlying idea, we can calculate the fuzzy entropy values for features by using similarity values between the ideal vectors and sample vectors which we want to classify. By summing the entropy values for all the samples in the training set for the feature, we get t entropy values for t features. Now if the uncertainty is high, we expect to get high entropy values and if similarities are high (or low) we except to get low entropy values. Now based on this underlying assumption next we find the feature which had the highest entropy value when similarities between ideal vector value of this feature and sample vector values of this feature was calculated. Next the decision

Table 3
Classification results with Pima-Indian diabetes data.

Method	Mean accuracy (%)	Variance	Dimension	AUROC	Selected features
Sim	75.29	0.0351	8	0.7620	All
Sim + F1	75.84	0.0061	7	0.7031	1,2,3,4,5,7,8
Sim + F2	75.97	0.0141	2	0.6668	1,2

Table 4
Classification results with Parkinsons data.

Method	Mean accuracy (%)	Variance	Dimension	AUROC	Selected features
Sim	79.22	0.0152	22	0.7620	All
Sim + F1	85.03	0.0053	2	0.8661	19 and 20
Sim + F2	84.52	0.0068	1	0.8688	19

Table 5
Classification results with breast cancer data.

Method	Mean accuracy (%)	Variance	Dimension	AUROC	Selected features
Sim	97.49	0.0350	9	0.9558	All
Sim + F1	97.10	0.0007	8	0.9705	1,2,3,4,5,6,7,8
Sim + F2	97.18	0.0004	5	0.9707	1,2,3,6,8

Table 6
Classification results with dermatology data.

Method	Mean accuracy (%)	Variance	Dimension	Removed features
Sim	96.04	0.0076	34	None
Sim + F1	98.15	0.0174	32	31,33
Sim + F2	98.28	0.0116	29	13,23,24,30 and 31

for removing this feature is made according to the highest entropy value since with this feature we make the assumption that it is not contributing much for the deviation between classes and most informative features are getting lowest entropy values. After removing this feature, the procedure can then be repeated and features can be removed more using this idea.

The feature selection method is presented in pseudo-code form in the following:

```
Require: $idealvec[1,\ldots,l]$, $Datalearn[1,\ldots,m]$
  for $j = 1$ to $m$ do
    for $i = 1$ to $t$ do
      for $k = 1$ to $l$ do
      $sim[j][i][k] = (1 - idealvec[j][i][k]^p - Datalearn[i][j]^p)^{(1/p)}$
      end for
    end for
  end for
  Sort similarity values $sim[i][j][k]$ according to feature set $U$
  for $i = 1$ to $t$ do
    $H[i] = -\sum_{x \in U} \mu_i[x] ln \mu_i(x) + (1 - \mu_i(x)) ln(1 - \mu_i(x))$
  end for
  $J = arg\ max_i H[i]$
  Remove $J$:th feature values from the data.
```

In the algorithm, we are having m samples, t features and l classes. After the similarities have been calculated, we have to sort the similarity values in matrix format again simply by tiling similarity matrices from each class to one larger similarity matrix of size $ml \times t$ from which the fuzzy entropy values for each feature summing through ml values can be calculated for each feature. After this, we simply find the feature with largest fuzzy entropy value and remove that feature from the data set.

3.2. Similarity classifier

The problem of classification is basically one of partitioning the attribute space into regions, one region for each category. Ideally, one would like to arrange this partitioning so that none of the decisions are ever wrong (Duda & Hart, 1973).

Consider we would like to classify a set X of objects into N different classes $C_1,\ldots,C_N$ by their attributes. We suppose that t is the number of different kinds of features $f_1,\ldots,f_t$ that we can measure from the objects. We suppose that the values for the magnitude of each attribute are normalized so that they can be presented as a value between $[0,1]$. Consequently, the objects we want to classify are vectors that belong to $[0,1]^t$.

First one must determine for each class the ideal vector $\mathbf{v}_i = (v_i(f_1),\ldots,v_i(f_t))$ that represents the class i as well as possible. This

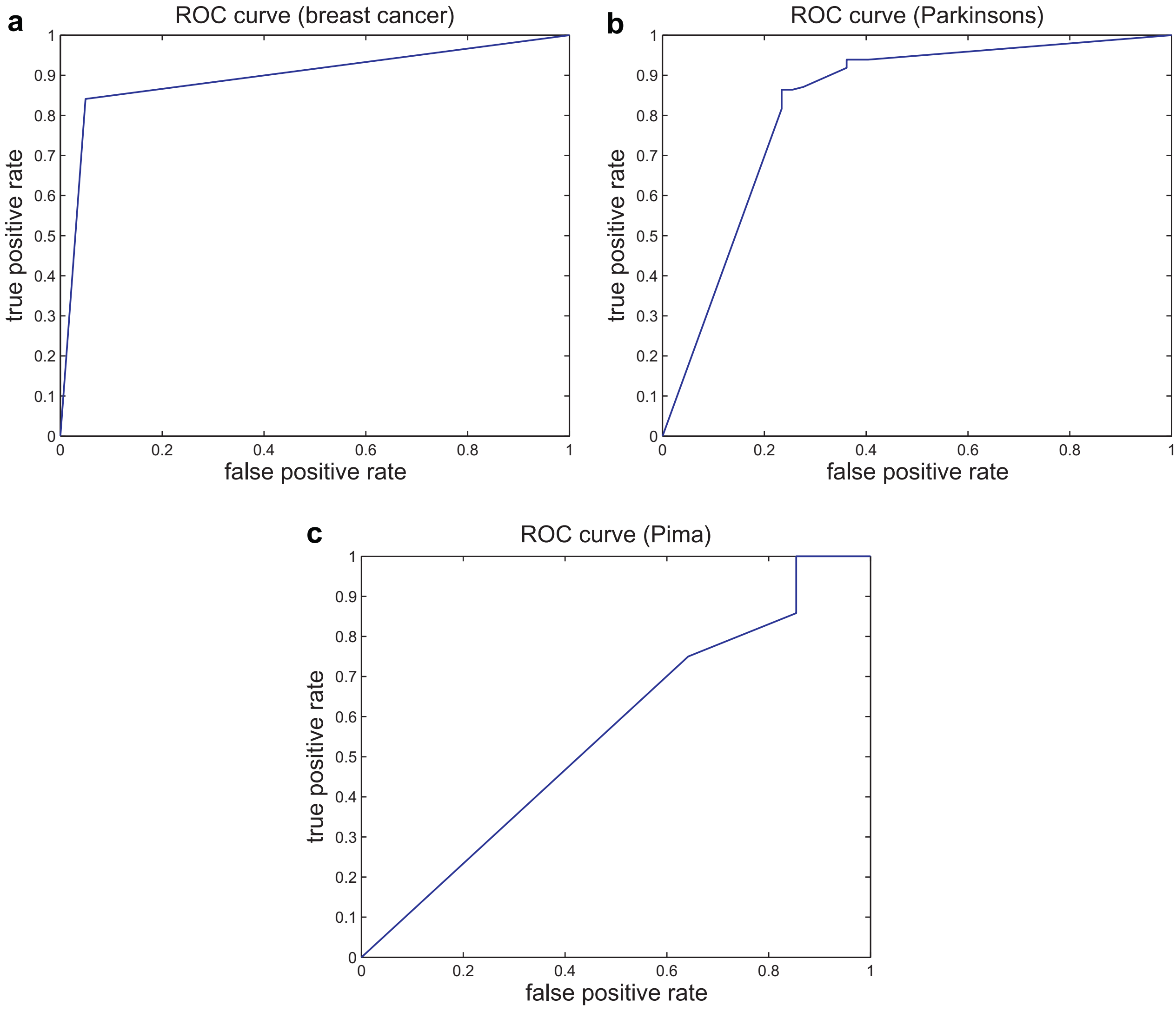

Fig. 1. ROC curves with data sets: (a) breast cancer, (b) Parkinsons and (c) Pima-Indian diabetes.

vector can be user defined or calculated from some sample set X_i of vectors $\mathbf{x} = (x(f_1),\ldots,x(f_t))$ which are known to belong to class C_i. We can use, e.g. the generalized mean for calculating $\mathbf{v}_i$, which is

$$v_i(r) = \left(\frac{1}{\sharp X_i}\sum_{\mathbf{x}\in X_i} x(f_r)^m\right)^{\frac{1}{m}}, \quad \forall r = 1,\ldots,t \tag{4}$$

where power value m (coming from the generalized mean) is fixed for all i, r and $\sharp X_i$ simply means the number of samples in class i.

Once the ideal vectors have been determined, then the decision to which class an arbitrarily chosen $\mathbf{x} \in X$ belongs is made by comparing it to each ideal vector. The comparison can be done, e.g. by using similarity in the generalized Łukasiewicz structure

$$S\langle \mathbf{x},\mathbf{v}\rangle = \left(\frac{1}{t}\sum_{r=1}^{t} w_r\left(1-|x(f_r)^p - v(f_r)^p|\right)^{m/p}\right)^{1/m} \tag{5}$$

for $\mathbf{x}, \mathbf{v} \in [0,1]^t$. Here, p is a parameter coming from the generalized Łukasiewicz structure (Luukka et al., 2001) (if $p = 1$ the equation again becomes its 'normal' form which holds in 'normal' Łukasiewicz structure or just simply a Łukasiewicz structure) and w_d is a weight parameter so that different weights can be given for different attributes to emphasize their importance if it seems appropriate. In this study, weights were set as one. The similarity measure has a strong mathematical background (Formato, Gerla, & Scarpati, 1999; Klawonn & Castro, 1995) and has proven to be a very efficient measure in classification (Luukka & Leppälampi, 2006). We decide that $\mathbf{x} \in C_i$ if

$$S\langle \mathbf{x},\mathbf{v}_i\rangle = \max_{i=1,\ldots,N} S\langle \mathbf{x},\mathbf{v}_i\rangle \tag{6}$$

In other words, the decision to which class the sample belongs is made according to which ideal vector the sample has the highest similarity value. There are several reasons why the Łukasiewicz structure is chosen in defining memberships of objects. One reason is that in the Łukasiewicz structure, it holds that the mean of many similarities is still a similarity (Turunen, 1999). Secondly, the Łukasiewicz structure has a strong connection to first-order logic (Novak, 1990) which is a well-studied area in modern mathematics. Thirdly, it also holds that any pseudo-metric induces similarity on a given non-empty set X with respect to the Łukasiewicz conjunction (Klawonn & Castro, 1995). Good sources of information about the Łukasiewicz structure can be found in Łukasiewicz (1970).

4. Classification results and comparison

In all data sets, data were split into half. One half was used for training and one half for testing. This procedure was repeated

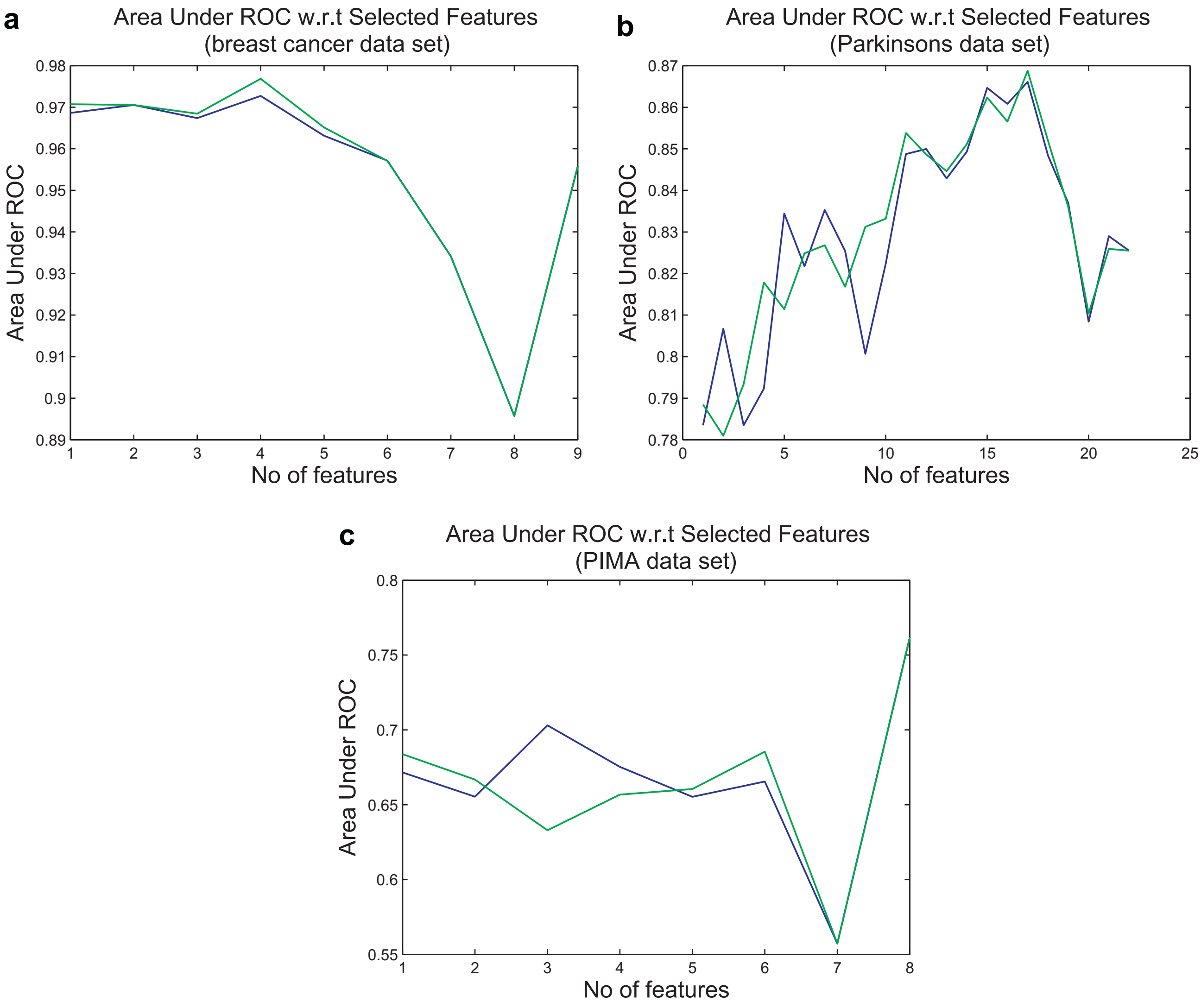

Fig. 2. Area under ROC curves w.r.t. reduced features for data sets: (a) breast cancer, (b) Parkinsons and (c) Pima-Indian diabetes.

randomly 30 times and mean classification accuracies and variances were computed. In Tables 3–6, results are reported for the data sets. In first column, the used method is given. First the results for just using similarity classifier is given (Sim), then results with similarity classifier and first feature extraction method are reported. In first feature extraction method De Luca and Termini (1971), suggestion for fuzzy entropy was used. Below this row again a combination with similarity classifier and feature extraction method is used, but now Parkash et al. (2008) suggestion for fuzzy entropy is used for feature extraction. Parkash et al. reported two ways to compute fuzzy entropy but results were the same for extracted features so only from one the results are given in Tables 3–6. Weights in the fuzzy entropy measures were set to one. Next in the result Tables 3–6 in second column, mean classification accuracies are reported, in third column variance and in fourth column, number of features used to get the highest mean classification accuracy is given. In fifth column for those data sets which are binary classification problems also the Area Under Receiver Operator Characteristic (AUROC) is computed and last the selected features for best classification performance is given. In Fig. 1 Receiver Operatior Characteristic (ROC) graphs are calculated for binary classification problem (data sets breast cancer, Parkinsons and Pima-Indian diabetes); there ROC graph is calculated for results where best mean classification accuracy was achieved. In Fig. 2, AUROCs are reported for reduced feature sets when similarity classifier is used. In Fig. 3, classification accuracies with reduced feature sets are reported for the two feature extraction methods used. In Fig. 4, classifiers mean accuracy is given with varying similarity parameters p and m. Next, we go into details of the results.

4.1. Pima-Indian diabetes

In Table 3 results are reported for Pima-Indian diabetes data set. Mean classification accuracy of 75.29% was achieved using similarity classifier. When first feature extraction method (F1) was used together with similarity classifier, a mean accuracy of 75.84% was achieved and with second feature extraction method (F2) a mean accuracy of 75.97% was gained. Mean accuracies with this data set are not much higher with feature extraction used but more noticeable is that with the highest mean accuracy (with F2), only first two features were needed. With this data set, one managed to reduce computational time quite significantly, and also simplify the model considerably. In Fig. 3(c), one can see that quite high mean classification accuracy was achieved using 2–8 features with both feature extraction methods. In Fig. 1(c), ROC graph is calculated for this data set and in Fig. 2(c), AUROCS can be seen w.r.t. selected features. From Fig. 4, classification accuracy w.r.t. similarity

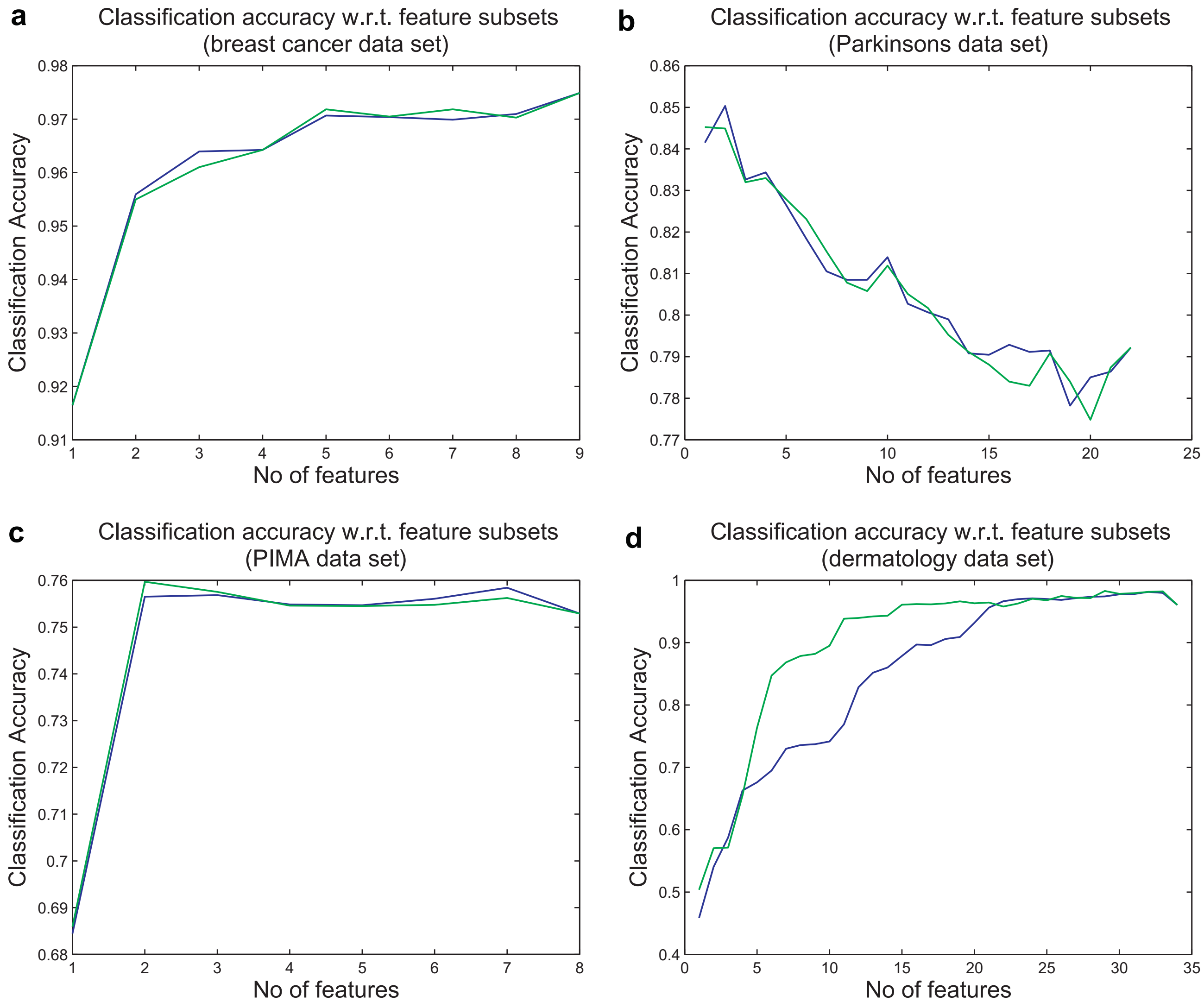

Fig. 3. Classification accuracies w.r.t. reduced features for data sets: (a) breast cancer, (b) Parkinsons, (c) Pima-Indian diabetes and (d) dermatology.

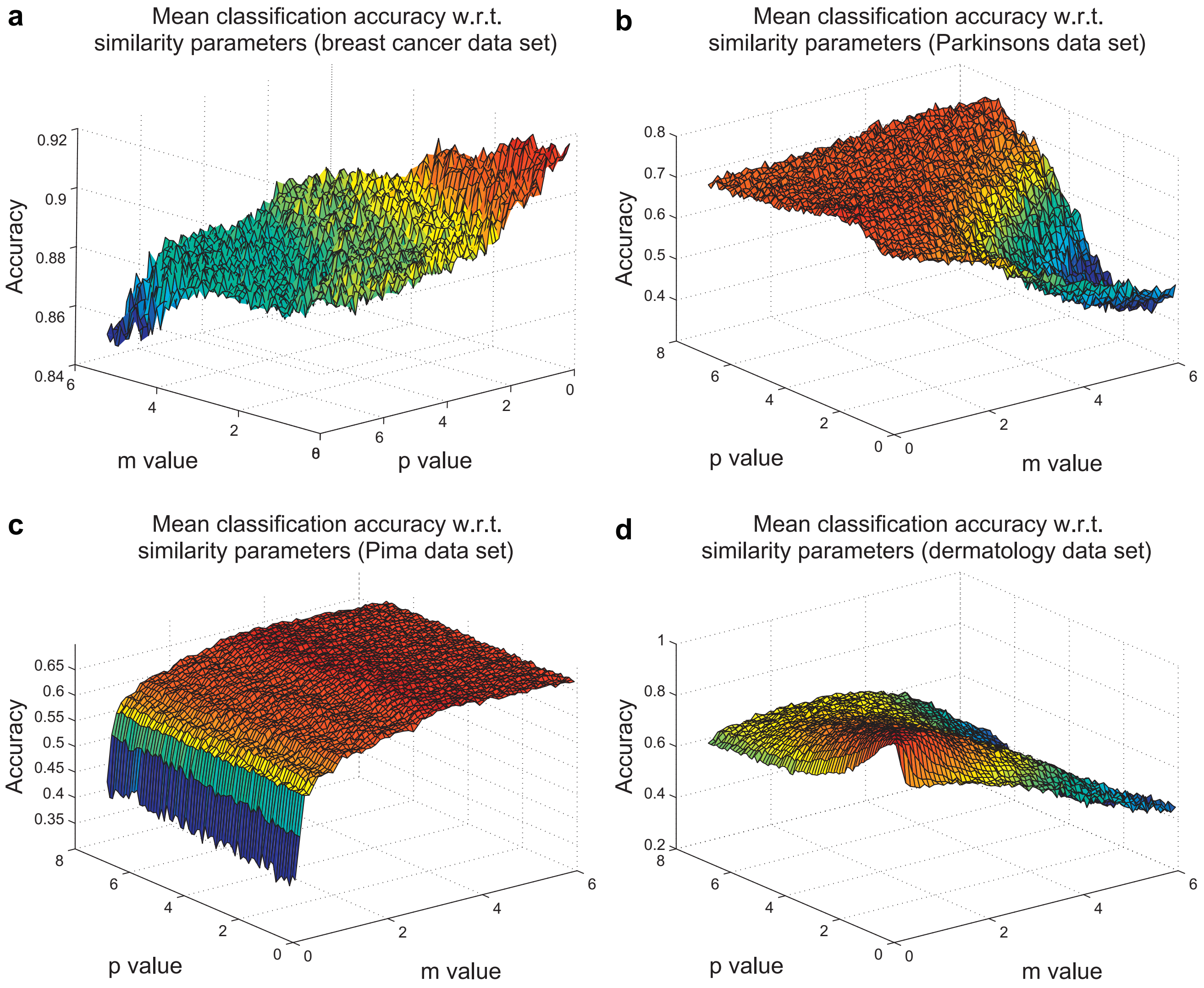

Fig. 4. Mean classification accuracies w.r.t. similarity parameters *p* and *m* for data sets: (a) breast cancer, (b) Parkinsons, (c) Pima-Indian diabetes and (d) dermatology.

parameter changes is given. As can be seen from the Figure as long as $m > 4$, quite good results were found.

4.2. Parkinsons data set

In Table 4, results from Parkinsons data set can be seen. Here by reducing the feature set, we managed to enhance the mean classification accuracy significantly. Improvement in mean classification accuracy was over 5% units and the best mean accuracy was 85.03% using similarity classifier and first feature extraction method. Moreover as can be seen from Table 4 and even better from Fig. 3, results were improving almost all the time, when features were reduced and finally best results were gained using only two features from the data set. This also suggests significant reduction in computational cost and simplifies the model a lot.

4.3. Breast cancer data set

In Table 5, one can see the results for breast cancer data set. There using similarity classifier without feature extraction the highest mean accuracy of 97.49% was gained. When feature extraction methods were used, one managed to get 97.10% and 97.18% mean classification accuracies and with second feature extraction method only five features was used thus also reducing computational cost. In Figs. 1 and 2, ROC graph and AUROCS w.r.t. reduced features can be seen. AUROCs actually where highest when only four features were used with this data set.

4.4. Dermatology data set

Results from the dermatology data set can be seen in Table 6. There since now we are dealing with six class classification problem, instead of two class classification problem ROC graph and AUROC are omitted. Also since data are 34 dimensional and best results were found with 29 dimensions we report removed features in the last column instead of listing all 29 features which were left in the model. Mean classification accuracy with using similarity classifier was 96.04% and both feature extraction methods managed to get higher classification accuracy with Sim + F1 98.15% mean classification accuracy and with Sim + F2, 98.28%, mean classification accuracy. Around 2% units, improvement was achieved with feature selection methods. Best subset of features consisted of 29 features so five features were managed to remove successfully.

5. Discussion

Using fuzzy entropy-based feature selection combined with similarity classifier, we managed to reduce the computational time and simplify the data set by using only subset of features instead of the whole data set to do the classification. Also, in all other cases except with breast cancer data set, also classification accuracy was managed to enhance. Using Pima-Indian diabetes and Parkinsons data, the required data set was managed to reduce to only two features where highest mean classification accuracy was found. This clearly makes the computational time much lower. It also simplifies the model so that it is not necessary to take so many measurements, but one can managed as well with just two features. This indicates that not so many measurements from patients are required to get to this accuracy. With Parkinsons data set, classification accuracy was also significantly enhanced when feature selection was performed. Mean classification accuracy of 85.03% with using just two features was achieved. Using feature extraction also managed to enhance classification accuracy with dermatology data set now with mean classification accuracy of 98.28% achived using 29 features from original 34. Experiments show that feature selection method using fuzzy entropy measures together with similarity classifier is giving good results.

References

Abdulla, W. H., & Kasabov, N. (2003). Reduced feature-set based parallel CHMM speech recognition systems. *Information Science, 156*, 21–38.

Aha, D. W., & Bankert, R. L. (1996). *A comparative evaluation of sequential feature selection algorithms. Learning from data: AI and statistics V*. Berlin: Springer.

Bandemer, H., & Näther, W. (1992). *Fuzzy data analysis*. Kluwer Academic Publisher.

Blum, A., & Langley, P. (1997). Selection of relevant features and examples in machine learning. *Artificial Intelligence, 97*, 245–271.

De Luca, A., & Termini, S. (1971). A definition of non-probabilistic entropy in setting of fuzzy set theory. *Information Control, 20*, 301–312.

Duda, R., & Hart, P. (1973). *Pattern classification and scene analysis*. John Wiley & Sons.

Formato, F., Gerla, G., & Scarpati, L. (1999). Fuzzy subgroups and similarities. *Soft Computing, 3*, 1–6.

Garret, D., Peterson, D. A., Anderson, C. W., & Thaut, M. H. (2003). Comparison of linear, nonlinear, and feature selection methods for EEG signal classification. *IEEE Transactions on Neural Systems and Rehabilitation Engineering, 11*, 141–144.

Gletsos, M., Mougiakakou, S. G., Matsopoulos, G. K., Nikita, K. S., Nikita, A. S., & Kelekis, D. (2003). A computer-aided diagnostic system to characterize CT focal liver lesions: design and optimization of a neural network classifier. *IEEE Transactions on Information Technology in Biomedicine, 7*, 153–162.

Hoffman, A. J., Hoogenboezem, C., Van der Merwe, N. T., & Tollig, C. J. A. (1998). Seismic buffer recognition using mutual information for selecting wavelet based features. In *IEEE international symposium on industrial electronics* (pp. 663–667).

Klawonn, F., & Castro, J. L. (1995). Similarity in fuzzy reasoning. *Mathware & Software Computing, 2*(3), 197–228.

Kohavi, R., & John, G. H. (1997). Wrappers for feature subset selection. *Artificial Intelligence, 7*, 273–323.

Kupinski, M. A., & Giger, M. L. (1997). Feature selsection and classifiers for the computerized detection of mass lesions in digital mammography. In *International conference on neural networks* (pp. 2460–2463).

Lee, W. L., Chen, Y. C., & Hsieh, K. S. (2003). Ultrasonic liver tissues classification by fractal feature vector based on M-band wavelet transform. *IEEE Transactions Medical Imagine, 22*, 382–392.

Lewin, D. I. (2000). Getting clinical about neural networks. *IEEE Intelligent Systems, 15*, 2–3.

Łukasiewicz, J. (1970). *Selected works*. Cambridge Univ. Press.

Luukka, P., Saastamoinen, K., & Könönen, V. (2001). A classifier based on the maximal fuzzy similarity in the generalized Łukasiewicz-structure. In *Proceedings of the FUZZ-IEEE 2001 conference, Melbourne, Australia*.

Luukka, P., & Leppälampi, T. (2006). Similarity classifier with generalized mean applied to medical data. *Computers in Biology and Medicine, 36*, 1026–1040.

Matsui, K., & Kosugi, Y. (1999). Image segmentation by neural-net classifiers with genetic selection of feature indices. In *IEEE international conference on image processing* (pp. 534–538).

McNitt-Gray, M. F., Huang, H. K., & Sayre, J. W. (1995). Feature selection in the pattern classification problem of digital chest radiograph segmentation. *IEEE Transactions on Medical Imagine, 14*, 537–547.

Newman, D. J., Hettich, S., Blake, C. L., & Merz, C. J. (2007). UCI Repository of machine learning databases. Irvine, CA: University of California, Department of Information and Computer Science. <http://www.ics.uci.edu/~mlearn/MLRepository.html>

Novak, V. (1990). On the syntactico-semantical completeness of first-order fuzzy logic. *Kybernetika*, 26.

Ozdemir, M., Embrechts, M. J., arciniegas, F., Breneman, C. M., Lockwood, L., & Bennett, K. P. (2001). Feature selection for in-silico drug design using genetic algorithms and neural networks. In *IEEE mountain workshop on soft computing in industrial applications* (pp. 53–57).

Parkash, O. M., Sharma, P. K., & Mahajan, R. (2008). New measures of weighted fuzzy entropy and their applications for the study of maximum weighted fuzzy entropy principle. *Information Sciences, 178*(11), 2389–2395.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 379–423, 623–659.

Turunen, E. (1999). *Mathematics behind fuzzy logic. Advances in soft computing*. Heidelberg: Physica-Verlag.

Übeyli, E. D., & Güler, I. (2005). Automatic detection of erythemato-squamous diseases using adaptive neuro-fuzzy inference systems. *Computers in Biology and Medicine, 35*(5), 147–165.

Yu, S., De Backer, S., & Scheunders, P. (2000). Genetic feature selection combined with composite fuzzy nearest neighbor classifiers for high-dimensional remote sensing data. In *IEEE international conference on systems, man, and cybernetics* (pp. 1912–1916).

Zadeh, L. (1965). Fuzzy sets. *Information and Control, 8*, 338–353.