Appl Intell (2008) 28: 69–82
DOI 10.1007/s10489-007-0042-6

Feature subset selection based on fuzzy entropy measures for handling classification problems

Jen-Da Shie · Shyi-Ming Chen

Received: 18 October 2006 / Accepted: 28 February 2007 / Published online: 18 May 2007

Abstract In this paper, we present a new method for dealing with feature subset selection based on fuzzy entropy measures for handling classification problems. First, we discretize numeric features to construct the membership function of each fuzzy set of a feature. Then, we select the feature subset based on the proposed fuzzy entropy measure focusing on boundary samples. The proposed method can select relevant features to get higher average classification accuracy rates than the ones selected by the MIFS method (Battiti, R. in IEEE Trans. Neural Netw. 5(4):537–550, 1994), the FQI method (De, R.K., et al. in Neural Netw. 12(10):1429–1455, 1999), the OFEI method, Dong-and-Kothari's method (Dong, M., Kothari, R. in Pattern Recognit. Lett. 24(9):1215–1225, 2003) and the OFFSS method (Tsang, E.C.C., et al. in IEEE Trans. Fuzzy Syst. 11(2):202–213, 2003).

Keywords Fuzzy entropy · Classification problems · Feature subset selection · Fuzzy logic · Membership grade

1 Introduction

In recent years, some feature subset selection methods have been proposed, such as similarity measures [26], gain-entropies [3], the relevance of features [1], the genetic algorithms method [4], the overall feature evaluation index (OFEI) [10], the feature quality index (FQI) [10], the mutual information-based feature selector (MIFS) [2], classifiability measures [12], neuro-fuzzy approaches [11, 20], ..., etc. In [12], Dong and Kothari pointed out that the task of feature subset selection aims to reduce the number of features used in classification or recognition tasks. It is obvious that a data set might have irrelevant and relevant features. If we can properly select relevant features to deal with classification problems, we can increase the classification accuracy rates [5–8].

In this paper, we present a new method for dealing with feature subset selection based on fuzzy entropy measures for handling classification problems. First, we discretize numeric features to construct the membership function of each fuzzy set of a feature. Then, we select the feature subset based on the proposed fuzzy entropy measure focusing on boundary samples. We use four different kinds of classifiers (i.e., LMT [17], Naive Bayes [15], SMO [21], and C4.5 [22]) to compare the average classification accuracy rates of the proposed feature subset selection method with the methods used to compare with the proposed method in the experiments, i.e., the OFFSS method [26], the OFEI method [10], the FQI method [10], the MIFS method [2] and Dong-and-Kothari's method [12], where the Iris data set, the Breast cancer data set, the Pima Diabetes data set, the MPG data set, the Cleve data set, the Correlated data set, the M of N-3-7-10 data set, the Crx data set, the Monk-1 data set, the Monk-2 data set and the Monk-3 data set are used in our experiments (Data Source: UCI Repository of Machine Learning Databases and Domain Theories, ftp://ftp.ics.uci.edu/pub/machine-learning-databases/). The proposed feature subset selection method can select features to get higher average classification accuracy rates than the ones selected by the MIFS method [2], the FQI method [10], the OFEI method [10], Dong-and-Kothari's method [12] and the OFFSS method [26].

J.-D. Shie · S.-M. Chen (✉)
Department of Computer Science and Information Engineering, National Taiwan University of Science and Technology, 43, Section 4, Keelung Road, Taipei 106, Taiwan
e-mail: smchen@mail.ntust.edu.tw

The rest of this paper is organized as follows. In Sect. 2, we briefly review some entropy measures [16, 18, 19, 24, 27] and propose a new method to calculate the fuzzy entropy of a fuzzy set. In Sect. 3, we present a new method to calculate the fuzzy entropy of a feature and propose an algorithm to construct the membership function of each fuzzy set of a feature. In Sect. 4, we present an algorithm for feature subset selection. In Sect. 5, we use the proposed feature subset selection algorithm to select feature subsets from different kinds of data sets. We also make some experiments to compare the average classification accuracy rate of the features selected by the proposed method with the ones selected by the MIFS method [2], the FQI method [10], the OFEI method [10], Dong-and-Kothari's method [12] and the OFFSS method [26] based on different kinds of classifiers. The conclusions are discussed in Sect. 6.

2 Fuzzy entropy measures

In this section, we briefly review the existing entropy measures [16, 18, 19, 24, 27] and propose a new method to calculate the fuzzy entropy of a fuzzy set.

The entropy measure is commonly used in information theory, where Shannon's entropy [24] is widely used. It can be used to characterize the impurity of a collection of samples. Let X be a discrete random variable with a finite set containing n elements, where $X = \{x_1, x_2, \ldots, x_n\}$. If an element x_i occurs with a probability $p(x_i)$, then the amount of information $I(x_i)$ associated with x_i is defined as follows:

$$I(x_i) = -\log_2 p(x_i). \tag{1}$$

The entropy $H(X)$ of X is defined as follows:

$$H(X) = -\sum_{i=1}^{n} p(x_i)\log_2 p(x_i), \tag{2}$$

where n denotes the number of elements and $p(x_i)$ denotes the occurring probability of the element x_i.

In [27], Zadeh defined a fuzzy entropy on a fuzzy set $\tilde{A}$ for a finite set $X = \{x_1, x_2, \ldots, x_n\}$ with respect to the probability distribution $P = \{p_1, p_2, \ldots, p_n\}$, shown as follows:

$$H = -\sum_{i=1}^{n} \mu_{\tilde{A}}(x_i) p_i \log p_i, \tag{3}$$

where $\mu_{\tilde{A}}$ denotes the membership function of $\tilde{A}$, $\mu_{\tilde{A}}(x_i)$ denotes the grade of membership of x_i belonging to the fuzzy set $\tilde{A}$, p_i denotes the probability of x_i, and $1 \le i \le n$.

In [19], Luca and Termini defined a fuzzy entropy measure based on Shannon's entropy [24]. They presented a set of axioms for a fuzzy entropy measure. The axioms of a fuzzy entropy measure are reviewed from [19] as follows. Assume that A is a fuzzy set defined in the universe of discourse X and μ_A is the membership function of the fuzzy set A, where $\mu_A(x): X \to [0, 1]$, $\mu_A(x)$ indicates the grade of membership of x belonging to the fuzzy set A, and $x \in X$. The axioms of a fuzzy entropy measure $H(A)$ of a fuzzy set A are as follows [19]:

Axiom 1: $H(A) = 0$ *iff* $A \in X$ is a crisp set.

Axiom 2: $H(A)$ is the maximum *iff* $\mu_A(x) = 0.5, \forall x \in A$.

Axiom 3: if $\tilde{A}$ is less fuzzy than $\tilde{B}$, then $H(\tilde{A}) \le H(\tilde{B})$.

Axiom 4: $H(A) = H(A^c)$, where $A^c = 1 - A$, i.e., A^c denotes the complement of A.

The fuzzy entropy of a fuzzy set proposed by Luca et al. is reviewed from [18] as follows:

$$H = -K\sum_{j=1}^{n}[(\mu_A(x_j)\log\mu_A(x_j)) + (1-\mu_A(x_j))\log(1-\mu_A(x_j))], \tag{4}$$

where μ_A denotes the membership function of the fuzzy set A, $\mu_A(x_j)$ denotes the grade of membership of x_j belonging to the fuzzy set A, $1 \le j \le n$ and $k = 1/n$.

In [16], Kosko used the concepts of overlap and underlap to define a fuzzy entropy $H(A)$ of a fuzzy set A based on the geometry of hypercube, shown as follows:

$$H(A) = \frac{\sum_{i=1}^{n}(\mu_A(x_i) \wedge \mu_A^C(x_i))}{\sum_{i=1}^{n}(\mu_A(x_i) \vee \mu_A^C(x_i))}, \tag{5}$$

where μ_A denotes the membership function of the fuzzy set A, $\mu_A(x_i)$ denotes the grade of membership of x_i belonging to the fuzzy set A, $\mu_A^C(x_i)$ denotes the complement of $\mu_A(x_i)$, $1 \le i \le n$, $\wedge$ denotes the minimum operator, and $\vee$ denotes the maximum operator.

In [18], Lee et al. presented a fuzzy entropy measure of an interval, based on Shannon's entropy measure [24] and Luca's axioms [19]. The fuzzy entropy measure proposed by Lee et al. is reviewed from [18] as follows. Assume that a set of samples R is divided into a set C of classes, and assume that a feature dimension is divided into I intervals. Let $\tilde{A}$ be a fuzzy set defined in a feature dimension, R_i be a subset of R distributed in the ith interval, and R_{ic} be a subset of R_i labeled as class c, where $c \in C$. The matching degree MD_c of the samples of class c in the ith interval belonging to the fuzzy set $\tilde{A}$, where $c \in C$, is defined as follows [18]:

$$MD_c(\tilde{A}) = \frac{\sum_{r\in R_{ic}} \mu_{\tilde{A}}(r)}{\sum_{r\in R_i} \mu_{\tilde{A}}(r)}. \tag{6}$$

The fuzzy entropy $IFEc(\tilde{A})$ of the samples of class c in the ith interval belonging to the fuzzy set $\tilde{A}$, where $c \in C$, is defined as follows:

$$IFE_c(\tilde{A}) = -MD_c(\tilde{A}) \log_2 MD_c(\tilde{A}). \quad (7)$$

The fuzzy entropy $IFE(\tilde{A})$ of the samples in the ith interval belonging to the fuzzy set $\tilde{A}$ is defined as follows:

$$IFE(\tilde{A}) = \sum_{c \in C} IFE_c(\tilde{A}). \quad (8)$$

The fuzzy entropy TFE_i of the ith interval in a feature dimension is defined as follows:

$$TFE_i = \sum_{v \in V_i} IFE(v), \quad (9)$$

where V_i denotes the set of fuzzy sets in the ith interval in a feature dimension.

In this paper, we present a new fuzzy entropy measure of a fuzzy set, shown as follows.

Definition 2.1 Assume that a set X of samples is divided into a set C of classes. The class degree $CD_c(\tilde{A})$ of the samples of class c, where $c \in C$, belonging to the fuzzy set $\tilde{A}$ is defined by:

$$CD_c(\tilde{A}) = \frac{\sum_{x \in X_c} \mu_{\tilde{A}}(x)}{\sum_{x \in X} \mu_{\tilde{A}}(x)}, \quad (10)$$

where X_c denotes the samples of class c, $c \in C$, $\mu_{\tilde{A}}$ denotes the membership function of the fuzzy set $\tilde{A}$, $\mu_{\tilde{A}}(x)$ denotes the membership grade of x belonging to the fuzzy set $\tilde{A}$, and $\mu_{\tilde{A}}(x) \in [0, 1]$.

Definition 2.2 The fuzzy entropy $FE_c(\tilde{A})$ of the samples of class c, where $c \in C$, belonging to the fuzzy set $\tilde{A}$ is defined as follows:

$$FE_c(\tilde{A}) = -CD_c(\tilde{A}) \log_2 CD_c(\tilde{A}). \quad (11)$$

Definition 2.3 The fuzzy entropy $FE(\tilde{A})$ of a fuzzy set $\tilde{A}$ is defined by:

$$FE(\tilde{A}) = \sum_{c \in C} FE_c(\tilde{A}). \quad (12)$$

Assume that there is a sample data set shown in Fig. 1, where the symbols "O" and "X" denote the positive samples and the negative samples, respectively. The corresponding fuzzy sets $\tilde{A}$, $\tilde{B}$ and $\tilde{C}$ of feature A are shown in Fig. 2. The numeric feature A is divided into three intervals I_1, I_2 and I_3 which correspond to the three fuzzy sets $\tilde{A}$, $\tilde{B}$ and $\tilde{C}$, respectively, where $I_1 = [0, 2]$, $I_2 = [2, 4]$ and $I_3 = [4, 6]$. The entropies of the intervals I_1 and I_2 calculated by Shannon's entropy measure [24] and the proposed fuzzy entropy measure are calculated as follows.

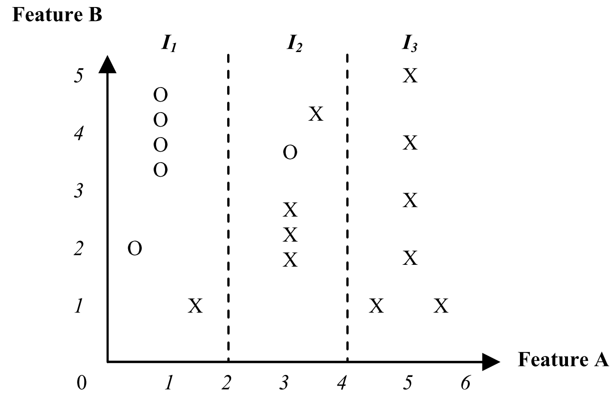

Fig. 1 The distribution of the samples with two features

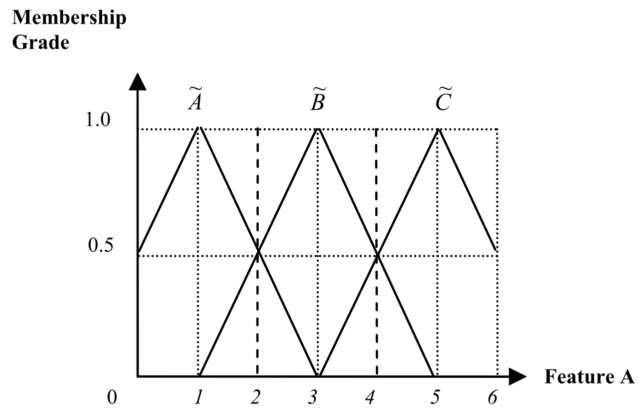

Fig. 2 The corresponding fuzzy sets of feature A

Based on Shannon's entropy measure, (i.e., (1–2)), we can calculate the entropies of the intervals I_1 and I_2, respectively, shown as follows:

$$H(I_1) = -(p(o)\log_2 p(o) + p(\times)\log_2 p(\times))$$
$$= -\left(\frac{5}{6} \times \log_2 \frac{5}{6} + \frac{1}{6} \times \log_2 \frac{1}{6}\right) \cong 0.65,$$
$$H(I_2) = -\left(\frac{1}{6} \times \log_2 \frac{1}{6} + \frac{5}{6} \times \log_2 \frac{5}{6}\right) \cong 0.65.$$

Based on the proposed method (i.e., (10–12)), we can calculate the fuzzy entropies of the fuzzy sets $\tilde{A}$ and $\tilde{B}$, respectively, shown as follows:

(1) Calculate the fuzzy entropy of the fuzzy set $\tilde{A}$:
(i) Calculate the summation of the membership grades of the samples of each class belonging to the fuzzy

set $\tilde{A}$:

$$\sum_{x\in X_o} \mu_{\tilde{A}}(x) = 0.75+1+1+1+1 = 4.75,$$

$$\sum_{x\in X_\times} \mu_{\tilde{A}}(x) = 0.75.$$

(ii) Based on (10), calculate the class degree of the samples of each class belonging to the fuzzy set $\tilde{A}$:

$$CD_o(\tilde{A}) = \frac{4.75}{4.75+0.75} = \frac{4.75}{5.5} = 0.864,$$

$$CD_\times(\tilde{A}) = \frac{0.75}{4.75+0.75} = \frac{0.75}{5.5} = 0.136.$$

(iii) Based on (11) and (12), the fuzzy entropy $FE(\tilde{A})$ of the fuzzy set $\tilde{A}$ is calculated as follows:

$$\begin{aligned} FE(\tilde{A}) &= FE_o(\tilde{A}) + FE_\times(\tilde{A}) \\ &= -(CD_o(\tilde{A})\log_2 CD_o(\tilde{A}) \\ &\quad + CD_\times(\tilde{A})\log_2 CD_\times(\tilde{A})) \\ &= -\left(\frac{4.75}{5.5}\times\log_2\frac{4.75}{5.5} + \frac{0.75}{5.5}\times\log_2\frac{0.75}{5.5}\right) \\ &\cong 0.575. \end{aligned}$$

(2) Calculate the fuzzy entropy of the fuzzy set $\tilde{B}$:

(i) Calculate the summation of the membership grades of the samples of each class belonging to the fuzzy set $\tilde{B}$:

$$\sum_{x\in X_o} \mu_{\tilde{B}}(x) = 1,$$

$$\sum_{x\in X_\times} \mu_{\tilde{B}}(x) = 0.25+1+1+1+0.75+0.75+0.25 = 5.$$

(ii) Based on (10), calculate the class degree of the samples of each class belonging to the fuzzy set $\tilde{B}$:

$$CD_o(\tilde{B}) = \frac{1}{1+5} = \frac{1}{6} = 0.167,$$

$$CD_\times(\tilde{B}) = \frac{5}{1+5} = \frac{5}{6} = 0.833.$$

(iii) Based on (11) and (12), the fuzzy entropy $FE(\tilde{B})$ of the fuzzy set $\tilde{B}$ is calculated as follows:

$$\begin{aligned} FE(\tilde{B}) &= FE_o(\tilde{B}) + FE_\times(\tilde{B}) \\ &= -(CD_o(\tilde{B})\log_2 CD_o(\tilde{B}) \\ &\quad + CD_\times(\tilde{B})\log_2 CD_\times(\tilde{B})) \\ &= -\left(\frac{1}{6}\times\log_2\frac{1}{6} + \frac{5}{6}\times\log_2\frac{5}{6}\right) \\ &\cong 0.65. \end{aligned}$$

From the above results, we can see that Shannon's entropy of the interval I_1 is equal to that of the interval I_2 (i.e., it can not distinguish the entropies of the intervals I_1 and I_2). But the proposed fuzzy entropy measure can indicate that the sample distribution in the interval I_2 is more ambiguous than that in the interval I_1.

3 The proposed fuzzy entropy measures of features

In this section, we present a fuzzy entropy measure of a feature and present an algorithm to construct the membership function of each fuzzy set of a feature. A feature can be described by several linguistic terms [29], where each linguistic term can be represented by a fuzzy set [27] characterized by a membership function. The proposed fuzzy entropy measure of a feature is defined as follows.

Definition 3.1 Fuzzy entropy $FFE(f)$ of a feature f is defined by:

$$FFE(f) = \sum_{v\in V} \frac{S_v}{S} FE(v), \tag{13}$$

where V denotes the set of fuzzy sets of feature f, $FE(v)$ denotes the fuzzy entropy of the fuzzy set v, S denotes the summation of the membership grades of the samples belonging to each fuzzy set of the feature f, and S_v denotes the summation of the membership grades of the samples belonging to the fuzzy set v.

There are two categories of features, where the one is nominal and the other one is numeric. Both of them have their corresponding membership functions of fuzzy sets. Each value of a nominal feature can be regarded as a fuzzy set, where its membership function is defined as follows:

$$\mu_u(x) = \begin{cases} 1, & \text{if } x = u, \\ 0, & \text{otherwise} \end{cases} \tag{14}$$

where $u \in U$, U denotes a set of values of a nominal feature, and μ_u denotes the membership function of the fuzzy set u. For example, the set of values of the feature "Sex" is {male, female}. When the value of the feature "Sex" is "male", the membership grades are: $\mu_{\text{male}}(\text{male}) = 1$ and $\mu_{\text{female}}(\text{male}) = 0$.

A numeric feature can be discretized into finite fuzzy sets. The number of fuzzy sets will affect the result of classification. Therefore, the discretization of a numeric feature is an important process. Using unsupervised learning techniques to discretize a numeric feature is a good method,

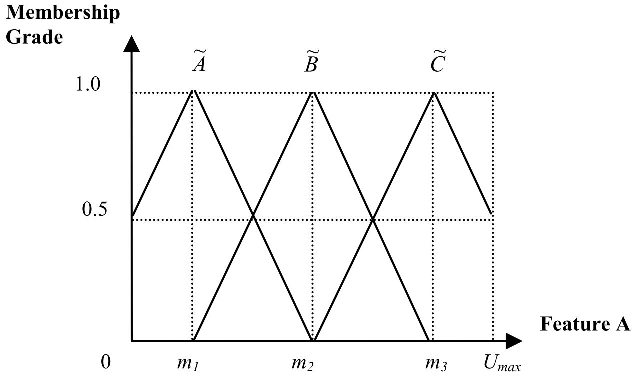

Fig. 3 A numeric feature **A** with fuzzy sets $\tilde{A}$, $\tilde{B}$ and $\tilde{C}$, where the clusters centers of $\tilde{A}$, $\tilde{B}$ and $\tilde{C}$, are m_1, m_2 and m_3, respectively

where the k-means clustering algorithm [14] is widely used. In this paper, we apply the k-means clustering algorithm to generate k cluster centers, where $k \geq 2$, and then construct their corresponding membership functions, where the cluster centers are used as the centers of fuzzy sets, respectively. Assume that m_1, m_2 and m_3 are the cluster centers of three clusters of a numeric feature **A**, respectively. Then, we can construct their corresponding membership functions of the fuzzy sets $\tilde{A}$, $\tilde{B}$ and $\tilde{C}$, respectively, as shown in Fig. 3, where $\mu_{\tilde{A}}(0) = 0.5$, $\mu_{\tilde{A}}(m_1) = 1$, $\mu_{\tilde{A}}(m_2) = 0$, $\mu_{\tilde{B}}(m_1) = 0$, $\mu_{\tilde{B}}(m_2) = 1$, $\mu_{\tilde{B}}(m_3) = 0$, $\mu_{\tilde{C}}(m_2) = 1$, $\mu_{\tilde{C}}(m_3) = 0$, and $\mu_{\tilde{C}}(U_{\max}) = 0.5$.

The fuzzy entropy of a feature decreases when the number of clusters increases. However, too many clusters could cause the overfitting problem [23] and reduce their classification accuracy rates when they classify new instances [22]. In this paper, we use a threshold value T_c to avoid the overfitting problem, where $T_c \in [0, 1]$. When the decreasing rate of the fuzzy entropy of a feature is less than the threshold value T_c given by the user, we stop increasing the number of clusters, where the decreasing rate of a fuzzy entropy of a feature is obtained by subtracting the fuzzy entropy of the feature calculated by clustering the values of the feature into k clusters from the fuzzy entropy of the feature calculated by clustering the values of the feature into $k-1$ clusters. In the following, we present an algorithm to construct the membership functions of the fuzzy sets of a numeric feature, shown as follows:

Step 1: Initially, set the number k of clusters to 2.

Step 2: Use the k-means clustering algorithm to generate k cluster centers based on the values of a feature, where $k \geq 2$, shown as follows:
/* assign initial values to the k clusters centers. */
for $i = 1$ to k do

$$m_i = x_{\frac{i}{k}};$$

repeat
{
/* assign each sample to the cluster which has the minimum Euclidean distance, where "$\arg\min_{k\in K} \|x - m_k\|^2$" returns one of such k that minimizes the equation $\|x - m_k\|^2$ and "$\| \bullet \|$" denotes the Euclidean norm. */
for all $x \in X$
{

$$i = \arg\min_{k\in K} \|x - m_k\|^2;$$

$$Cluster_i = Cluster_i \cup \{x\}$$

};
/* calculate a new cluster center m_i for each cluster, where n_i denotes the number of items in the ith cluster and $1 \leq i \leq k$. */
for $i = 1$ to k do

$$m_i = \frac{\sum_{x\in Cluster_i} x}{n_i};$$

} until each cluster is not changed.

Step 3: Construct the membership functions of the fuzzy sets based on these k cluster centers, respectively, shown as follows:
/* assign neighbor cluster centers to the ith cluster center "m_i", where "m_L" denotes the left "cluster center" of m_i, "m_R" denotes the right "cluster center" of m_i, "$U_{\min}$" denotes the minimum value of a feature, and "$U_{\max}$" denotes the maximum value of a feature. */

$$\text{let } m_L = \begin{cases} U_{\min} - (m_i - U_{\min}), & \text{if } i = 1, \\ m_{i-1}, & \text{otherwise}; \end{cases}$$

$$\text{let } m_R = \begin{cases} U_{\max} + (U_{\max} - m_i), & \text{if } i = K, \\ m_{i+1}, & \text{otherwise}; \end{cases}$$

/* construct the membership function μ_{v_i} of the fuzzy set v_i based on the ith cluster center m_i, where "Max" denotes maximum operator. */

$$\text{let } \mu_{v_i}(x) = \begin{cases} Max\{1 - \frac{m_i - x}{m_i - m_L}, 0\}, & \text{if } x \leq m_i, \\ Max\{1 - \frac{x - m_i}{m_R - m_i}, 0\}, & \text{if } x > m_i. \end{cases}$$

Step 4: Based on (4–7), calculate the fuzzy entropy of feature f, shown as follows:
for $i = 1$ to k do

$$FE(v_i) = \sum_{c\in C} FE_c(v_i);$$

$$\text{let } FFE(f) = \sum_{v\in V} \frac{s_v}{s} FE(v).$$

Step 5: If the decreasing rate of the fuzzy entropy of feature f is larger than the threshold value T_c given by the user, where $T_c \in [0, 1]$, then let
$k = k + 1$ and go to **Step 2**. Otherwise, let $k = k - 1$ and **Stop**.

4 The proposed feature subset selection algorithm

In this section, we present a new method for feature subset selection. The proposed method uses "boundary samples" instead of a full set of samples to select the feature subset. First, we introduce the concept of "boundary samples". Then, we define the fuzzy entropy of a feature subset. Finally, we propose a new algorithm for feature subset selection based on boundary samples.

The feature subset selection problem can be regarded as a dimension reduction problem [9, 13]. Assume that there is a two-dimensional feature space as shown in Fig. 4, where the symbols "O" and "X" denote the positive samples and the negative samples, respectively. We can reduce Fig. 4 into two one-dimensional feature spaces as shown in Fig. 5. Dimension reduction will increase the entropy of data because some information will be omitted at the same time. Thus, we should avoid the decrease of classification accuracy caused by omitting important features.

In a dimension reduction problem [13], each feature might have incorrectly classified samples. Thus, an optimal feature subset is a set of correlated features [15]. It means that the samples incorrectly classified by a feature could be correctly classified by other features. "Boundary samples" are incorrectly classified samples of features, and we should focus on them for feature subset selection. For example, Table 1 shows an example data set with three nominal features, where the samples incorrectly classified by feature **A** are Sample 1, Sample 2 and Sample 5 due to the fact that the classes of these samples with the same feature value are ambiguous. Thus, the value of feature **A** with incorrectly classified samples is "black". In the same way, the samples shown in Table 1 incorrectly classified by feature **B** are Sample 2, Sample 5 and Sample 6. Thus, we can only use Sample 2 and Sample 5 to calculate the entropy of the feature subset {**A**, **B**}. Because Sample 1 can be correctly classified by feature **B**, it can also be correctly classified by the feature subset {**A**, **B**}. Thus, Sample 1 can be omitted. In the same way, because Sample 3 and Sample 4 can be correctly classified by feature **A** or feature **B** and Sample 6 can be correctly classified by feature **A**, Sample 3, Sample 4 and Sample 6 can also be correctly classified by the feature subset {**A**, **B**}. Thus, Sample 3, Sample 4 and Sample 6 can be omitted, too. Therefore, we can reduce the number of samples from 6 to 2, i.e., Sample 2 and Sample 5.

A feature subset can be regarded as a collection of features. For example, in Table 1, the values of the feature subset {**A**, **B**} are {(black, ocean), (black, lake), (black, river), (white, ocean), (white, lake), (white, river), (red, ocean), (red, lake), (red, river)}. In Table 1, Sample 2 and Sample 5 are called the "boundary samples" due to the fact that when the values of feature **A** and feature **B** of Sample 2 and Sample 5 are "black" and "lake", respectively, they get the different labels "positive" and "negative", respectively. Thus, we can calculate the entropy of the feature subset {**A**, **B**} by only using the boundary samples, i.e., Sample 2 and Sample 5. However**,** we can not use the boundary samples to calculate the fuzzy entropy of a feature subset directly. We

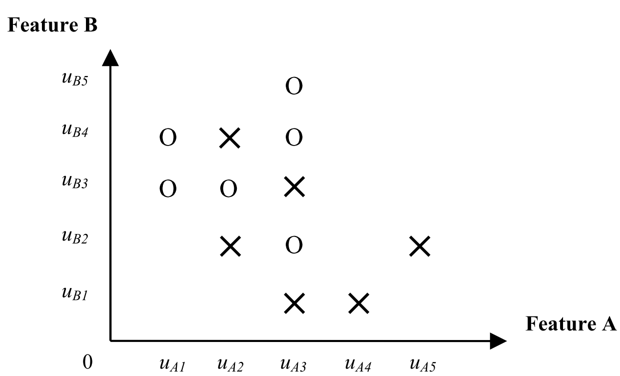

Fig. 4 A two-dimensional feature space with two classes

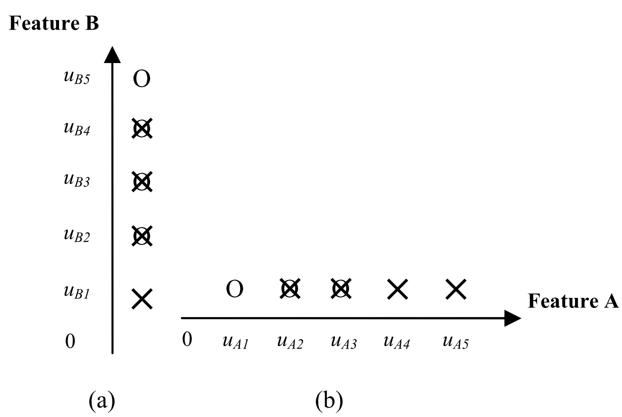

Fig. 5 Two one-dimensional feature spaces. (a) Feature **A** is omitted, (b) Feature **B** is omitted

Table 1 An example of data set

Sample No.	Feature **A**	Feature **B**	Feature **C**	Classes
1	black	ocean	summer	positive
2	black	lake	winter	positive
3	white	ocean	fall	positive
4	red	river	winter	negative
5	black	lake	fall	negative
6	red	lake	fall	negative

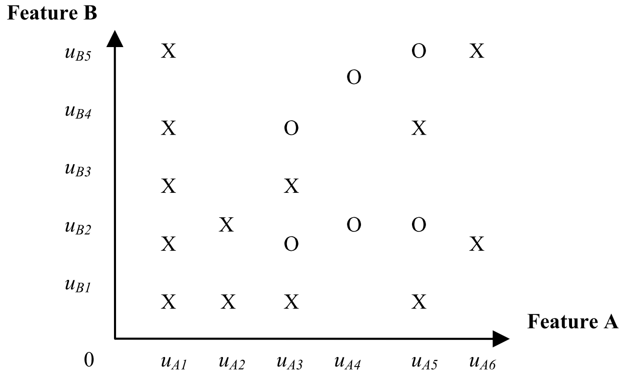

Fig. 6 The samples distribution with two numeric features and two classes

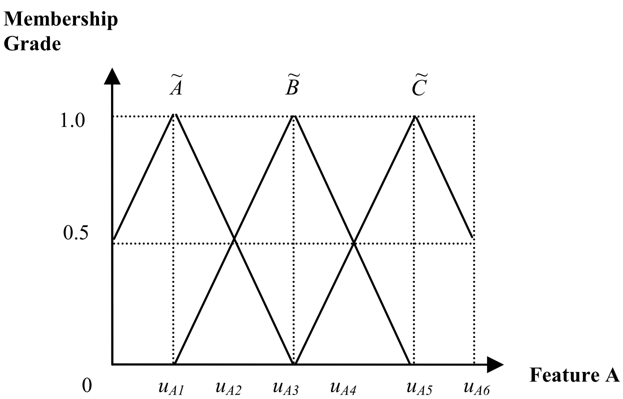

Fig. 7 The corresponding fuzzy sets of the numeric feature **A**

can use an indirect method to simplify the feature subset selection process described as follows.

Assume that there is a sample data with two numeric features shown in Fig. 6, where the symbols "O" and "X" denote the positive samples and the negative samples, respectively. The corresponding fuzzy sets of the numeric feature A are shown in Fig. 7.

The fuzzy entropies of the fuzzy sets $\tilde{A}$, $\tilde{B}$ and $\tilde{C}$ can be calculated by (10–12), shown as follows:

(1) Calculate the fuzzy entropy of the fuzzy set $\tilde{A}$:

(i) Calculate the summation of the membership grades of the samples of each class belonging to the fuzzy set $\tilde{A}$:

$$\sum_{x \in X_o} \mu_{\tilde{A}}(x) = 0,$$

$$\sum_{x \in X_\times} \mu_{\tilde{A}}(x) = 1 + 1 + 1 + 1 + 1 + 0.5 + 0.5 = 6.$$

(ii) Based on (10), calculate the class degree of the samples of each class belonging to the fuzzy set $\tilde{A}$:

$$CD_o(\tilde{A}) = \frac{0}{0+6} = \frac{0}{6} = 0,$$

$$CD_\times(\tilde{A}) = \frac{6}{0+6} = \frac{6}{6} = 1.$$

(iii) Based on (11) and (12), the fuzzy entropy $FE(\tilde{A})$ of the fuzzy set $\tilde{A}$ is calculated as follows:

$$\begin{aligned} FE(\tilde{A}) &= FE_o(\tilde{A}) + FE_\times(\tilde{A}) \\ &= -(CD_o(\tilde{A}) \log_2 CD_o(\tilde{A}) \\ &\quad + CD_\times(\tilde{A}) \log_2 CD_\times(\tilde{A})) \\ &= -(0 \times \log_2 0 + 1 \times \log_2 1) = 0. \end{aligned}$$

(2) Calculate the fuzzy entropy of the fuzzy set $\tilde{B}$:

(i) Calculate the summation of the membership grades of the samples of each class belonging to the fuzzy set $\tilde{B}$:

$$\sum_{x \in X_o} \mu_{\tilde{B}}(x) = 1 + 1 + 0.5 + 0.5 = 3,$$

$$\sum_{x \in X_\times} \mu_{\tilde{B}}(x) = 0.5 + 0.5 + 1 + 1 = 3.$$

(ii) Based on (10), calculate the class degree of the samples of each class belonging to the fuzzy set $\tilde{B}$:

$$CD_o(\tilde{B}) = \frac{3}{3+3} = \frac{3}{6} = 0.5,$$

$$CD_\times(\tilde{B}) = \frac{3}{3+3} = \frac{3}{6} = 0.5.$$

(iii) Based on (11) and (12), the fuzzy entropy $FE(\tilde{B})$ of the fuzzy set $\tilde{B}$ is calculated as follows:

$$\begin{aligned} FE(\tilde{B}) &= FE_o(\tilde{B}) + FE_\times(\tilde{B}) \\ &= -(CD_o(\tilde{B}) \log_2 CD_o(\tilde{B}) \\ &\quad + CD_\times(\tilde{B}) \log_2 CD_\times(\tilde{B})) \\ &= -\left(\frac{3}{6} \times \log_2 \frac{3}{6} + \frac{3}{6} \times \log_2 \frac{3}{6}\right) = 1. \end{aligned}$$

(3) Calculate the fuzzy entropy of the fuzzy set $\tilde{C}$:

(i) Calculate the summation of the membership grades of the samples of each class belonging to the fuzzy set $\tilde{C}$:

$$\sum_{x \in X_o} \mu_{\tilde{C}}(x) = 0.5 + 0.5 + 1 + 1 = 3,$$

$$\sum_{x \in X_\times} \mu_{\tilde{C}}(x) = 1 + 1 + 0.5 + 0.5 = 3.$$

(ii) Based on (10), calculate the class degree of the samples of each class belonging to the fuzzy set $\tilde{C}$:

$$CD_o(\tilde{C}) = \frac{3}{3+3} = \frac{3}{6} = 0.5,$$

$$CD_\times(\tilde{C}) = \frac{3}{3+3} = \frac{3}{6} = 0.5.$$

(iii) Based on (11) and (12), the fuzzy entropy $FE(\tilde{C})$ of the fuzzy set $\tilde{C}$ is calculated as follows:

$$\begin{aligned}FE(\tilde{C}) &= FE_o(\tilde{C}) + FE_\times(\tilde{C})\\ &= -(CD_o(\tilde{C})\log_2 CD_o(\tilde{C})\\ &\quad + CD_\times(\tilde{C})\log_2 CD_\times(\tilde{C}))\\ &= -\left(\frac{3}{6}\times\log_2\frac{3}{6}+\frac{3}{6}\times\log_2\frac{3}{6}\right)\\ &= 1.\end{aligned}$$

From Fig. 6, we can see that the samples whose values of feature **A** are smaller than u_{A3} can be correctly classified by feature **A**. If we omit the samples whose values of feature **A** are smaller than u_{A3}, then it will affect the fuzzy entropy of the fuzzy set $\tilde{B}$. Therefore, we must use a indirect method to calculate the fuzzy entropy of a feature subset by focusing on boundary samples. While we calculate the fuzzy entropy of a feature subset, we omit the fuzzy sets having lower fuzzy entropies. Thus, in the previous example, we can omit the fuzzy set $\tilde{A}$ to calculate the fuzzy entropy of the feature subset $\{\mathbf{A}, \mathbf{B}\}$.

In this paper, we use a threshold value T_r, where $T_r \in [0, 1]$, to omit the fuzzy sets of a feature whose maximum class degree is larger than or equal to the threshold value T_r given by the user for feature subset selection. According to Definition 2.1, we can see that there are n "class degrees" of a set of samples belonging to a fuzzy set with respect to n classes, respectively. The maximum class degree of a fuzzy set is defined as the maximum among these n "class degrees". If the maximum class degree of a fuzzy set is larger than or equal to the given threshold value T_r, where $T_r \in [0, 1]$, then the fuzzy set will be omitted to reduce the number of fuzzy sets of the feature. Then, we can construct the extension matrix of the membership grades of the values of a feature subset. Before we do this, we have to construct the extension matrices of all the features. The extension matrix of the membership grades of the values of a feature belonging to the fuzzy sets of this feature is defined as follows.

Definition 4.1 The extension matrix EM_f of the membership grades of the values of a feature f belonging to fuzzy sets of this feature is defined as follows:

$$EM_f = \begin{bmatrix} \mu_{v_1}(r_{1f}) & \cdots & \mu_{v_m}(r_{1f}) \\ \vdots & \vdots & \vdots \\ \mu_{v_1}(r_{nf}) & \cdots & \mu_{v_m}(r_{nf}) \end{bmatrix}_{n\times m}, \tag{15}$$

where n denotes the number of samples, m denotes the number of fuzzy sets of the feature f, $\mu_{v_z}(r_{pf})$ denotes the membership grade of the value r_{pf} of the feature f of the sample r_p belonging to the fuzzy set v_z, $1 \le p \le n$, and $1 \le z \le m$.

Let $EM_f[g, h]$ denote the element at row g and column h of an extension matrix EM_f, where $1 \le g \le n$, n denotes the number of samples, $1 \le h \le m$, and m denotes the number of fuzzy sets of a feature f. According to Definition 4.1, we can see that the membership grade $\mu_{v_z}(r_{pf})$ of the value r_{pf} of the feature f of the sample r_p belonging to the fuzzy set v_z is stored at row p and column z of an extension matrix EM_f (i.e., $EM_f[p, z]$). Then, the class degree $CD_c(v)$ of a set of samples can be calculated from the extension matrix EM_f of membership grades of the values of a feature f, defined as follows.

Definition 4.2 The class degree $CD_c(v)$ of the samples of class c belonging to the fuzzy set v, is defined as follows:

$$CD_c(v) = \frac{\sum_{r\in R_c} EM_f[|r|, |v|]}{\sum_{r\in R} EM_f[|r|, |v|]}, \tag{16}$$

where R denotes a set of samples, R_c denotes the samples of class c in R, $|r|$ denotes the number of the sample r, $1 \le |r| \le n$, n denotes the number of samples, $|v|$ denotes the number of the fuzzy set v, $1 \le |v| \le m$, and m denotes the number of fuzzy sets of the feature f. (Note: The pth sample is mapped into the pth row of the extension matrix EM_f and the zth fuzzy set of the feature f is mapped into the zth column of the extension matrix EM_f.)

The fuzzy entropy $FFE(f)$ of a feature f can be calculated by Definition 2.2, Definition 2.3, Definition 3.1 and Definition 4.2, shown as follows:

$$FFE(f) = \sum_{v\in V}\left[\frac{s_v}{s}\times\sum_{c\in C}(-CD_c(v)\log_2 CD_c(v))\right]. \tag{17}$$

In the following, we propose a "combined-extension-matrix function" for constructing the extension matrix of the membership grades of the values of a feature subset. Assume that there are a set of samples with two features f_1 and f_2, n denotes the number of samples, T_r denotes a maximum class degree threshold value given by the user, where $T_r \in [0, 1]$, i denotes the number of fuzzy sets of the feature f_1 whose maximum class degree is smaller than the given threshold value T_r, j denotes the number of fuzzy sets of the feature f_2 whose maximum class degree is smaller than the given threshold value T_r, $\mu_{v_{1x}}(r_{pf_1})$ denotes the membership grade of the value r_{pf_1} of the feature f_1 of the sample r_p belonging to a fuzzy set v_{1x} of the feature f_1, where $1 \le x \le i$, and $\mu_{v_{2y}}(r_{pf_2})$ denotes the membership grade of the value r_{pf_2} of the feature f_2 of the sample r_p belonging to a fuzzy set v_{2y} of the feature f_2, where $1 \le y \le j$. Let "$\mu_{v_{1x}}(r_{pf_1}) \wedge \mu_{v_{2y}}(r_{pf_2})$" denote the membership grade of the values of the feature subset $\{f_1, f_2\}$ of the sample r_p belonging to the combined fuzzy set "$v_{1x\wedge 2y}$" of the feature subset $\{f_1, f_2\}$, where $\wedge$ denotes the minimum operator. The proposed "combined-extension-matrix function" is shown as follows:

Definition 4.3 The combined extension matrix function $CEM(f_1, f_2, T_r)$ for constructing the extension matrix of the membership grades of the values of a feature subset $\{f_1, f_2\}$ belonging to the combined fuzzy sets of this feature subset according to the maximum class degree threshold value T_r given by the user, where $T_r \in [0, 1]$, is defined by:

$$CEM(f_1, f_2, T_r) = \begin{bmatrix} \mu_{v_{11}}(r_{1f_1}) \wedge \mu_{v_{21}}(r_{1f_2}) & \cdots & \mu_{v_{11}}(r_{1f_1}) \wedge \mu_{v_{2j}}(r_{1f_2}) & \cdots & \mu_{v_{1i}}(r_{1f_1}) \wedge \mu_{v_{21}}(r_{1f_2}) & \cdots & \mu_{v_{1i}}(r_{1f_1}) \wedge \mu_{v_{2j}}(r_{1f_2}) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \mu_{v_{11}}(r_{nf_1}) \wedge \mu_{v_{21}}(r_{nf_2}) & \cdots & \mu_{v_{11}}(r_{nf_1}) \wedge \mu_{v_{2j}}(r_{nf_2}) & \cdots & \mu_{v_{1i}}(r_{nf_1}) \wedge \mu_{v_{21}}(r_{nf_2}) & \cdots & \mu_{v_{1i}}(r_{nf_1}) \wedge \mu_{v_{2j}}(r_{nf_2}) \end{bmatrix}_{n \times ij}. \quad (18)$$

Based on the extension matrix of the membership grades of the values of a feature subset belonging to the combined fuzzy sets of this feature subset, the class degree of the samples of a class belonging to a combined fuzzy set of a feature subset can be calculated by (16). Then, the fuzzy entropy of the samples of a class belonging to a combined fuzzy set of a feature subset and the fuzzy entropy of a combined fuzzy set of a feature subset can be calculated by (11) and (12), respectively. Then, we propose a fuzzy entropy measure of a feature subset focusing on boundary samples, shown as follows.

Definition 4.4 The fuzzy entropy measure $BSFFE(f_1, f_2)$ of a feature subset $\{f_1, f_2\}$ focusing on boundary samples is defined as follows:

$$BSFFE(f_1, f_2) = \begin{cases} \frac{S_{1B}}{S_1} \times \sum_{w \in V_{FS}} \frac{S_w}{S_{FS}} FE(w) \\ \quad + \sum_{v_1 \in V_{1UB}} \frac{S_{v_1}}{S_1} FE(v_1), \\ \quad \text{if } \frac{S_{1B}}{S_1} < \frac{S_{2B}}{S_2}, \\ \frac{S_{2B}}{S_2} \times \sum_{w \in V_{FS}} \frac{S_w}{S_{FS}} FE(w) \\ \quad + \sum_{v_2 \in V_{2UB}} \frac{S_{v_2}}{S_2} FE(v_2), \\ \quad \text{otherwise} \end{cases} \quad (19)$$

where S_1 denotes the summation of the membership grades of the values of the feature f_1 of the samples belonging to each fuzzy set of the feature f_1, S_{1B} denotes the summation of the membership grades of the values of the feature f_1 of the samples belonging to the fuzzy sets of the feature f_1 whose maximum class degree is smaller than the threshold value T_r given by the user, where $T_r \in [0, 1]$, V_{FS} denotes the set of combined fuzzy sets of the feature subset $\{f_1, f_2\}$, S_{FS} denotes the summation of the membership grades of the values of the feature subset $\{f_1, f_2\}$ of the samples belonging to each combined fuzzy set of the feature subset $\{f_1, f_2\}$, S_w denotes the summation of the membership grades of the values of the feature subset $\{f_1, f_2\}$ of the samples belonging to a combined fuzzy set w, $FE(w)$ denotes the fuzzy entropy of a combined fuzzy set w, V_{1UB} denotes the set of fuzzy sets of the feature f_1 whose maximum class degree is larger than or equal to the threshold value T_r, S_{v_1} denotes the summation of the membership grades of the values of the feature f_1 of the samples belonging to a fuzzy set v_1 of the feature f_1, and $FE(v_1)$ denotes the fuzzy entropy of a fuzzy set v_1 of the feature f_1. Moreover, S_2 denotes the summation of the membership grades of the values of the feature f_2 of the samples belonging to the fuzzy sets of the feature f_2, S_{2B} denotes the summation of the membership grades of the values of the feature f_2 of the samples belonging to the fuzzy sets of the feature f_2 whose maximum class degree is smaller than the threshold value T_r given by the user, where $T_r \in [0, 1]$, V_{2UB} denotes the set of fuzzy sets of the feature f_2 whose maximum class degree is larger than or equal to the threshold value T_r, S_{v_2} denotes the summation of the membership grades of the values of the feature f_2 of the samples belonging to a fuzzy set v_2 of the feature f_2, and $FE(v_2)$ denotes the fuzzy entropy of a fuzzy set v_2 of the feature f_2.

Assume that a set R of samples is divided into a set C of classes, where $R = \{r_1, r_2, \ldots, r_n\}$, F denotes a set of candidate features and FS denotes the selected feature subset. The proposed algorithm for feature subset selection is now presented as follows:

Step 1: /* Construct the extension matrix EM_f of the membership grades of the values of each feature f belonging to fuzzy sets of each feature f and calculate the fuzzy entropy $FFE(f)$ of each feature f, respectively. */
for each $f \in F$ do
{
Based on (15), construct the extension matrix EM_f of the membership grades of the values of the feature f belonging to the fuzzy sets of the feature f,

shown as follows:

$$EM_f = \begin{bmatrix} \mu_{v_1}(r_{1\,f}) \cdots \mu_{v_m}(r_{1\,f}) \\ \vdots \quad \vdots \quad \vdots \\ \mu_{v_1}(r_{nf}) \cdots \mu_{v_m}(r_{n\,f}) \end{bmatrix}_{n\times m};$$

based on (16), calculate the class degree $CD_c(v)$ of the samples of each class c belonging to each fuzzy set v of the feature f, where $c \in C$;
based on (11) and (12), calculate the fuzzy entropy $FE(v)$ of each fuzzy set v of the feature f;
based on (13), calculate the fuzzy entropy $FFE(f)$ of the feature f
}.

Step 2: /* Put the feature with the minimum fuzzy entropy into the selected feature subset FS and remove it from the set F of candidate features. */
let $\hat{f} = \arg\min_{f\in F} FFE(f)$, where the symbol "$\arg\min_{f\in F} FFE(f)$" returns one of such a feature f that minimizes the function $FFE(f)$.
let $E_{FS} = FFE(\hat{f})$;
let $FS = \{\hat{f}\}$;
let $F = F - \{\hat{f}\}$.

Step 3: /* Repeatedly put the feature which can reduce the fuzzy entropy of the feature subset into FS until no such a feature exists. */
repeat
{
for each $f \in F$ do
{
based on (18), construct the extension matrix $EM_{FS\cup\{f\}}$ of membership grades of the values of the feature subset $FS\cup\{f\}$ according to the maximum class degree threshold value T_r given by the user, where $T_r \in [0, 1]$, shown as follows: $EM_{FS\cup\{f\}} = CEM(FS, f, T_r)$;

based on (16), calculate the class degree $CD_c(v)$ of the samples of each class c belonging to each combined fuzzy set v of the feature subset $FS\cup\{f\}$, where $c \in C$; based on (11) and (12), calculate the fuzzy entropy $FE(v)$ of each combined fuzzy set v of the feature subset $FS\cup\{f\}$;
based on (19), calculate the fuzzy entropy $BSFFE(FS, f)$ of the feature subset $FS\cup\{f\}$ focusing on boundary samples
};
let $\hat{f} = \arg\min_{f\in F} BSFFE(FS, f)$, where the symbol "$\arg\min_{f\in F} BSFFE(FS, f)$" returns one of such a feature f that minimizes the function $BSFFE(FS, f)$;
let $D = E_{FS} - BSFFE(FS, \hat{f})$;
let $E_{FS} = BSFFE(FS, \hat{f})$;
let $FS = FS \cup \{\hat{f}\}$;
let $F = F - \{\hat{f}\}$
} until ($E_{FS} = 0$ *or* $D \le 0$ *or* $F = \phi$);
let FS be the selected feature subset.

5 Experimental results

We have implemented the proposed method by using IBM Lotus Notes Version 4.6 (http://www-306.ibm.com/software/lotus/) on a Pentium 4 PC and have made two experiments, where four different kinds of classifiers (i.e., LMT [17], Naive Bayes [15], SMO [21], and C4.5 [22]) are used in the experiments. The first experiment uses four different kinds of UCI data sets (ftp://ftp.ics.uci.edu/pub/machine-learning-databases/), i.e., the Iris data set, the breast cancer data set, the Pima diabetes data set, and the MPG data set, for comparing the average classification accuracy rate of the features selected by the proposed method with the ones selected by the OFFSS method [26], the OFEI method [10], the FQI method [10] and the MIFS method [2], respectively. The second experiment uses eight different kinds of UCI data sets (ftp://ftp.ics.uci.edu/pub/machine-learning-databases/), i.e., the Pima diabetes data set, the Cleve data set, the Correlated data set, the M of N-3-7-10 data set, the Crx data set, the Monk-1 data set, the Monk-2 data set and the Monk-3 data set, for comparing the average classification accuracy rate of the features selected by the proposed method with the ones selected by the method presented in [12]. These two experiments are discussed as follows:

(1) The First Experiment: The Iris data set, the Breast cancer data set, the Pima Diabetes data set, and the MPG data set are used in this experiment. First, we apply the proposed method to select feature subsets of these four data sets (i.e., the Iris data set, the Breast cancer data set, the Pima Diabetes data set and the MPG data set), respectively. The proposed method consists of two major steps. The first step defines the corresponding membership function of each fuzzy set of each feature. The second step select feature subsets based on the proposed fuzzy entropy measure focusing on

Table 2 The threshold value *Tc* and *Tr* used in the proposed method

Data sets	The threshold value *Tc*	The threshold value *Tr*
Iris data set	0.2	0.9
Breast cancer data set	0.1	0.9
Pima diabetes data set	0.2	0.75
MPG data set	0.03	0.6

Table 3 A comparison of feature subsets selected by different methods

Data sets	Feature subsets selected by different methods				
	OFFSS	OFEI	FQI	MIFS	The proposed method
Iris data set	{4, 3}	{4, 3}	{4, 3}	{4, 3}	{4, 3}
Breast cancer data set	{6, 3, 1, 2}	{6, 1, 3, 2}	{6, 1, 8, 3}	{6, 3, 2, 7}	{6, 2, 1, 8, 5, 3}
Pima diabetes data set	{2, 6, 7}	{2, 3, 6}	{8, 2, 1}	{2, 6, 8}	{2, 6, 8, 7}
MPG data set	{6, 2, 5, 4}	{4, 5, 6, 2}	{4, 6, 3, 2}	{4, 6, 2, 1}	{4, 6, 3}

Table 4 A comparison of the average classification accuracy rates of different methods

Data sets	Classifiers	Average classification accuracy rates of different methods				
		OFFSS	OFEI	FQI	MIFS	The proposed method
Iris data set	LMT	$94.67 \pm 4.27\%$	$94.67 \pm 4.27\%$	$94.67 \pm 4.27\%$	$94.67 \pm 4.27\%$	$94.67 \pm 4.27\%$
	Naive Bayes	$96.00 \pm 4.00\%$	$96.00 \pm 4.00\%$	$96.00 \pm 4.00\%$	$96.00 \pm 4.00\%$	$96.00 \pm 4.00\%$
	SMO	$96.00 \pm 4.00\%$	$96.00 \pm 4.00\%$	$96.00 \pm 4.00\%$	$96.00 \pm 4.00\%$	$96.00 \pm 4.00\%$
	C4.5	$96.00 \pm 5.33\%$	$96.00 \pm 5.33\%$	$96.00 \pm 5.33\%$	$96.00 \pm 5.33\%$	$96.00 \pm 5.33\%$
Breast cancer data set	LMT	$95.90 \pm 2.15\%$	$95.90 \pm 2.15\%$	$96.49 \pm 2.09\%$	$95.46 \pm 1.79\%$	$96.49 \pm 2.08\%$
	Naive Bayes	$96.19 \pm 2.56\%$	$96.19 \pm 2.56\%$	$96.49 \pm 1.88\%$	$95.31 \pm 1.58\%$	$96.63 \pm 1.97\%$
	SMO	$96.34 \pm 2.19\%$	$96.34 \pm 2.19\%$	$97.07 \pm 1.85\%$	$96.05 \pm 2.62\%$	$97.07 \pm 2.27\%$
	C4.5	$95.61 \pm 2.70\%$	$95.61 \pm 2.70\%$	$96.93 \pm 1.90\%$	$95.16 \pm 2.86\%$	$96.02 \pm 2.57\%$
Pima diabetes data set	LMT	$76.83 \pm 3.79\%$	$76.04 \pm 3.63\%$	$73.56 \pm 4.68\%$	$75.53 \pm 4.39\%$	$77.22 \pm 4.52\%$
	Naive Bayes	$76.57 \pm 3.65\%$	$76.83 \pm 4.36\%$	$74.09 \pm 5.43\%$	$76.44 \pm 5.50\%$	$77.47 \pm 4.93\%$
	SMO	$75.91 \pm 4.96\%$	$75.91 \pm 3.80\%$	$75.39 \pm 4.93\%$	$75.91 \pm 4.97\%$	$77.08 \pm 5.06\%$
	C4.5	$75.01 \pm 3.72\%$	$74.36 \pm 4.27\%$	$71.74 \pm 3.18\%$	$74.61 \pm 4.86\%$	$74.88 \pm 5.89\%$
MPG data set	LMT	$81.13 \pm 5.67\%$	$81.13 \pm 5.67\%$	$82.38 \pm 7.28\%$	$84.17 \pm 7.26\%$	$81.87 \pm 6.74\%$
	Naive Bayes	$78.31 \pm 7.63\%$	$78.31 \pm 7.63\%$	$79.59 \pm 6.79\%$	$76.28 \pm 8.25\%$	$80.60 \pm 7.01\%$
	SMO	$80.58 \pm 7.21\%$	$80.58 \pm 7.21\%$	$81.61 \pm 6.99\%$	$76.77 \pm 4.12\%$	$81.86 \pm 8.25\%$
	C4.5	$79.83 \pm 7.84\%$	$79.83 \pm 7.84\%$	$79.58 \pm 8.24\%$	$81.37 \pm 9.05\%$	$79.93 \pm 7.78\%$

Note: All results are reported as mean ± standard deviation computed from 10 independent trials

Table 5 The threshold values Tc and Tr used in the proposed method

Data sets	The threshold value Tc	The threshold value Tr
Pima diabetes data set	0.2	0.75
Cleve data set	0.001	0.8
Correlated data set	N/A	0.95
M of N-3-7-10 data set	N/A	0.9
Crx data set	0.001	0.7
Monk-1 data set	N/A	0.9
Monk-2 data set	N/A	0.6
Monk-3 data set	N/A	0.95

Note: Because the features of the Correlated Data Set, the M of N-3-7-10 data set, the Monk-1 data set, the Monk-2 data set and the Monk-3 data set are nominal, the threshold Values Tc of these five data sets are not applied, denoted by the symbol "N/A"

boundary samples. The threshold value Tc used in the proposed algorithm for constructing the membership functions of the fuzzy sets of a numeric feature and the threshold value Tr used in the proposed algorithm for feature subset selection is shown in Table 2. A comparison of the experimental results of the feature subset selection for different methods is shown in Table 3.

Then, we use four different kinds of classifiers (i.e., LMT [17], Naive Bayes [15], SMO [21], and C4.5 [22]) to evaluate the performance of the selected feature subsets by different methods. We make the experiment in the environment of the free software Weka (http://www.cs.waikato.ac.nz/ml/weka/) on a Pentium 4 PC, where we use Weka to select different kinds of classifiers and different data sets with respect to the selected features by different methods. We apply the 10-fold cross-validation to the four data sets to get the average classification accuracy rates of different feature selection methods with respect to different classifiers as shown in Table 4. In the 10-fold cross-validation, we divide each data set into 10 subsets of approximately equal size and execute 10 times. Each time we select one of the

Table 6 A comparison of feature subsets selected by Dong-and-Kothari's method and the proposed method

Data sets	Feature subsets selected by different methods	
	Dong-and-Kothari's method	The proposed method
Pima diabetes data set	{2, 8, 1}	{2, 6, 8, 7}
Cleve data set	{10, 13, 12, 3, 9}	{13, 3, 12, 11, 1, 10, 2, 5, 6}
Correlated data set	{6, 1, 2, 3, 4}	{6, 1, 2, 3, 4}
M of N-3-7-10 data set	{4, 9, 5, 8, 3, 6, 7}	{4, 9, 8, 5, 3, 6, 7}
Crx data set	{8, 9, 13, 10}	{9}
Monk-1 data set	{5, 1, 2}	{5, 1, 2}
Monk-2 data set	{3, 6, 1, 2, 4, 5}	{5}
Monk-3 data set	{2, 5, 4, 1}	{5, 2, 4}

Table 7 A comparison of the average classification accuracy rates of Dong-and-Kothari's method with the proposed method

Data sets	Classifiers	Average classification accuracy rates of different methods	
		Dong-and-Kothari's method	The proposed method
Pima diabetes data set	LMT	$73.56 \pm 4.68\%$	$77.22 \pm 4.52\%$
	Naive Bayes	$73.43 \pm 1.57\%$	$77.47 \pm 4.93\%$
	SMO	$75.39 \pm 4.93\%$	$77.08 \pm 5.06\%$
	C4.5	$71.74 \pm 3.18\%$	$74.88 \pm 5.89\%$
Cleve data set	LMT	$83.17 \pm 4.24\%$	$82.87 \pm 6.23\%$
	Naive Bayes	$84.17 \pm 1.82\%$	$84.48 \pm 3.93\%$
	SMO	$84.47 \pm 5.59\%$	$83.51 \pm 6.09\%$
	C4.5	$76.90 \pm 8.71\%$	$76.90 \pm 8.40\%$
Correlated data set	LMT	$100.00 \pm 0.00\%$	$100.00 \pm 0.00\%$
	Naive Bayes	$86.03 \pm 3.75\%$	$86.03 \pm 3.75\%$
	SMO	$89.87 \pm 6.88\%$	$89.87 \pm 6.88\%$
	C4.5	$94.62 \pm 4.54\%$	$94.62 \pm 4.54\%$
M of N-3-7-10 data set	LMT	$100.00 \pm 0.00\%$	$100.00 \pm 0.00\%$
	Naive Bayes	$89.33 \pm 1.56\%$	$89.33 \pm 1.56\%$
	SMO	$100.00 \pm 0.00\%$	$100.00 \pm 0.00\%$
	C4.5	$100.00 \pm 0.00\%$	$100.00 \pm 0.00\%$
Crx data set	LMT	$85.22 \pm 4.04\%$	$85.22 \pm 4.04\%$
	Naive Bayes	$84.06 \pm 1.33\%$	$85.51 \pm 4.25\%$
	SMO	$85.80 \pm 3.71\%$	$85.80 \pm 3.71\%$
	C4.5	$85.36 \pm 4.12\%$	$85.51 \pm 4.25\%$
Monk-1 data set	LMT	$100.00 \pm 0.00\%$	$100.00 \pm 0.00\%$
	Naive Bayes	$74.97 \pm 1.95\%$	$74.97 \pm 1.95\%$
	SMO	$75.02 \pm 5.66\%$	$75.02 \pm 5.66\%$
	C4.5	$100.00 \pm 0.00\%$	$100.00 \pm 0.00\%$
Monk-2 data set	LMT	$67.36 \pm 1.17\%$	$67.36 \pm 1.17\%$
	Naive Bayes	$66.22 \pm 2.80\%$	$67.14 \pm 0.61\%$
	SMO	$67.14 \pm 0.61\%$	$67.14 \pm 0.61\%$
	C4.5	$67.14 \pm 0.61\%$	$67.14 \pm 0.61\%$
Monk-3 data set	LMT	$99.77 \pm 0.10\%$	$99.77 \pm 0.10\%$
	Naive Bayes	$97.22 \pm 0.47\%$	$97.21 \pm 2.71\%$
	SMO	$100.00 \pm 0.00\%$	$100.00 \pm 0.00\%$
	C4.5	$100.00 \pm 0.00\%$	$100.00 \pm 0.00\%$

Note: All results are reported as mean $\pm$ standard deviation computed from 10 independent trials

10 subsets as the testing data set and train the classifier by the remaining 9 subsets to get the classification accuracy rate with respect to each selected feature subset. After executing 10 times, we can get the average classification accuracy rate. From Table 4, we can see that the proposed method can select features to get higher average classification accuracy rates than the ones selected by the OFFSS method [26], the OFEI method [10], the FQI method [10] and the MIFS method [2].

(2) The Second Experiment: The Pima Diabetes data set, the Cleve data set, the Correlated data set, the M of N-3-7-10 data set, the Crx data set, the Monk-1 data set, the Monk-2 data set and the Monk-3 data set are used in this experiment. We apply the proposed method to select feature subsets from these eight data sets (i.e., the Pima diabetes data set, the Cleve data set, the Correlated data set, the M of N-3-7-10 data set, the Crx data set, the Monk-1 data set, the Monk-2 data set and the Monk-3 data set), respectively. The threshold value *Tc* used in the proposed algorithm for constructing the membership functions of the fuzzy sets of a numeric feature and the threshold value *Tr* used in the proposed algorithm for feature subset selection are shown in Table 5. A comparison of the results of the feature subset selection of Dong-and-Kothari's method [12] and the proposed method is shown in Table 6.

We use four different kinds of classifiers (i.e., LMT [17], Naive Bayes [15], SMO [21], and C4.5 [22]) to compare the average classification accuracy rates based on the features selected by the method proposed by Dong and Kothari [12] and the proposed method. We make the experiment in the environment of the free software Weka (http://www.cs.waikato.ac.nz/ml/weka/) on a Pentium 4 PC and apply the 10-fold cross-validation to the eight data sets to get the average classification accuracy rates as shown in Table 7. From Table 7, we can see that the proposed method can select features to get higher average classification accuracy rates than the ones selected by Dong-and-Kothari's method [12].

6 Conclusions

In this paper, we have presented a new method for feature subset selection based on the proposed fuzzy entropy measure for handling classification problems. The proposed method can deal with both numeric and nominal features. From the experimental results shown in Table 4 and Table 7, we can see that the proposed method can select relevant features to get higher average classification accuracy rates than the ones selected by the OFFSS method [26], the OFEI method [10], the FQI method [10], the MIFS method [2] and Dong-and-Kothari's method [12] with respect to different kinds of classifiers. In this paper, we use the k-means clustering algorithm to discrete the numeric features. In the future, we will investigate the effect of feature selection if other discretization methods are used.

Acknowledgements This work was supported in part by the National Science Council, Republic of China, under Grant NSC 93-2213-E-011-018.

References

1. Baim PW (1988) A method for attribute selection in inductive learning systems. IEEE Trans Pattern Anal Mach Intell 10(6):888–896
2. Battiti R (1994) Using mutual information for selecting features in supervised neural net learning. IEEE Trans Neural Netw 5(4):537–550
3. Caruana R, Freitag D (1994) Greedy attribute selection. In: Proceedings of international conference on machine learning, New Brunswick, NJ, pp 28–36
4. Chaikla N, Qi Y (1999) Genetic algorithms in feature selection. In: Proceedings of the 1999 IEEE international conference on systems, man, and cybernetics, Tokyo, Japan, vol 5, pp 538–540
5. Chen SM, Chang CH (2005) A new method to construct membership functions and generate weighted fuzzy rules from training instances. Cybern Syst 36(4):397–414
6. Chen SM, Chen YC (2002) Automatically constructing membership functions and generating fuzzy rules using genetic algorithms. Cybern Syst 33(8):841–862
7. Chen SM, Kao CH, Yu CH (2002) Generating fuzzy rules from training data containing noise for handling classification problems. Cybern Syst 33(7):723–748
8. Chen SM, Shie JD (2005) A new method for feature subset selection for handling classification problems. In: Proceedings of the 2005 IEEE international conference on fuzzy systems, Reno, NV, pp 183-188
9. Chen SM (1988) A new approach to handling fuzzy decision-making problems. IEEE Trans Syst Man Cybern 18(6):1012–1016
10. De RK, Basak J, Pal SK (1999) Neuro-fuzzy feature evaluation with theoretical analysis. Neural Netw 12(10):1429–1455
11. De RK, Pal NR, Pal SK (1997) Feature analysis: neural network and fuzzy set theoretic approaches. Pattern Recognit 30(10):1579–1590
12. Dong M, Kothari R (2003) Feature subset selection using a new definition of classifiability. Pattern Recognit Lett 24(9):1215–1225
13. Fodor IK, Kamath C (2002) Dimension reduction techniques and the classification of bent double galaxies. Comput Stat Data Anal 41(1):91–122
14. Hartigan JA, Wong MA (1979) A k-means clustering algorithm. J Roy Stat Soc Ser C 28(1):100–108
15. John GH, Kohavi R, Pfleger K (1994) Irrelevant features and the subset selection problem. In: Proceedings of the eleventh international conference on machine learning, San Francisco, CA, pp 121–129
16. Kosko B (1986) Fuzzy entropy and conditioning. Inf Scie 40(2):165–174
17. Landwehr N, Hall M, Frank E (2003) Logistic model trees. In: Proceedings of the 14th European conference on machine learning, Cavtat-Dubrovnik, Croatia, pp 241–252
18. Lee HM, Chen CM, Chen JM, Jou YL (2001) An efficient fuzzy classifier with feature selection based on fuzzy entropy. IEEE Trans Syst Man Cybern Part B Cybern 31(3):426–432

19. Luca AD, Termini S (1972) A definition of a non-probabilistic entropy in the setting of fuzzy set theory. Inf Control 20(4):301–312
20. Platt JC (1999) Using analytic QP and sparseness to speed training of support vector machines. In: Proceedings of the thirteenth annual conference on neural information processing systems, Denver, CO, pp 557–563
21. Pal SK, De RK, Bsask J (2000) Unsupervised feature selection: a neuro-fuzzy approach. IEEE Trans Neural Netw 11(2):366–376
22. Quinlan JR (1993) C4.5: programs for machine learning. Kaufmann, San Francisco
23. Schaffer C (1993) Overfitting avoidance as bias. Mach Learn 10(2):153–178
24. Shannon CE (1948) A mathematical theory of communication. Bell Syst Techn J 27(3):379–423
25. Shie JD, Chen SM (2006) A new approach for handling classification problems based on fuzzy information gain measures. In: Proceedings of the 2006 IEEE international conference on fuzzy systems, Vancouver, BC, Canada, pp 5427–5434
26. Tsang ECC, Yeung DS, Wang XZ (2003) OFFSS: optimal fuzzy-valued feature subset selection. IEEE Trans Fuzzy Syst 11(2):202–213
27. Zadeh LA (1965) Fuzzy sets. Inf Control 8:338–353
28. Zadeh LA (1965) Probability measures of fuzzy events. J Math Anal Appl 23(2):421–427
29. Zadeh LA (1975) The concept of linguistic variable and its application to approximate reasoning, I. Inf Sci 8(3):199–249

Jen-Da Shie received the B.S. degree from the Department of Information Management, National Kaohsiung First University of Science and Technology, Kaohsiung, Taiwan, Republic of China, in June 2004. He received the M.S. degree from the Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan, in October 2005. His research interests include fuzzy classification systems, data mining and artificial intelligence.

Shyi-Ming Chen is a Professor of the Department of Computer Science and Information Engineering, National Taiwan University of Science and Technology, Taipei, Taiwan, R. O. C. He received the Ph.D. degree in Electrical Engineering from National Taiwan University, Taipei, Taiwan, in June 1991. He has published more than 250 papers in referred journals, conference proceedings and book chapters. His research interests include fuzzy systems, information retrieval, knowledge-based systems, artificial intelligence, neural networks, data mining, and genetic algorithms. He has received several honors and awards, including the 1994 Outstanding Paper Award of the Journal of Information and Education, the 1995 Outstanding Paper Award of the Computer Society of the Republic of China, the Best Paper Award of the 1999 National Computer Symposium, Republic of China, the 1999 Outstanding Paper Award of the Computer Society of the Republic of China, the 2001 Outstanding Talented Person Award, Republic of China, for the contributions in Information Technology, the Outstanding Electrical Engineering Professor Award granted by the Chinese Institute of Electrical Engineering (CIEE), Republic of China, the 2003 Outstanding Paper Award of the Technological and Vocational Education Society, Republic of China, and the 2006 Outstanding Paper Award of the 11th Conference on Artificial Intelligence and Applications.

Dr. Chen is currently the President of the Taiwanese Association for Artificial Intelligence (TAAI). He is an Associate Editor of the IEEE Transactions on Systems, Man, and Cybernetics, Part C, an Associate Editor of the IEEE Computational Intelligence Magazine, an Associate Editor of the International Journal of Applied Intelligence, an Associate Editor of the Journal of Intelligent & Fuzzy Systems, an Associate Editor of the International Journal of Artificial Intelligence Tools, an Editor of the New Mathematics and Natural Computation Journal, an Associate Editor of the International Journal of Fuzzy Systems, an Editorial Board Member of the International Journal of Information and Communication Technology, an Editorial Board Member of the WSEAS Transactions on Systems, an Editor of the Journal of Advanced Computational Intelligence and Intelligent Informatics, an Associate Editor of the WSEAS Transactions on Computers, an Editorial Board Member of the International Journal of Computational Intelligence and Applications, an Editorial Board Member of the Advances in Fuzzy Sets and Systems Journal, an Editor of the International Journal of Soft Computing, an Editor of the Asian Journal of Information Technology, an Editorial Board Member of the International Journal of Intelligence Systems Technologies and Applications, an Editor of the Asian Journal of Information Management, an Associate Editor of the International Journal of Innovative Computing, Information and Control, an Editorial Board Member of the International Journal of Computer Applications in Technology, an Associate Editor of the Journal of Uncertain Systems, and an Editorial Board Member of the Advances in Computer Sciences and Engineering Journal. He was an Editor of the Journal of the Chinese Grey System Association from 1998 to 2003. He is listed in *International Who's Who of Professionals, Marquis Who's Who in the World, and Marquis Who's Who in Science and Engineering*. He is an IET Fellow.