Non-unique decision differential entropy-based feature selection

Qu, Yanpeng; Li, Rong; Deng, Ansheng ; Shang, Changjing; Shen, Qiang

Published in:
Neurocomputing

DOI:
10.1016/j.neucom.2018.10.112

Publication date:
2020

Citation for published version (APA):
Qu, Y., Li, R., Deng, A., Shang, C., & Shen, Q. (2020). Non-unique decision differential entropy-based feature selection. *Neurocomputing*, *393*, 187-193. https://doi.org/10.1016/j.neucom.2018.10.112

Non-unique Decision Differential Entropy-based Feature Selection

Yanpeng Qu[a,b,*], Rong Li[a], Ansheng Deng[a], Changjing Shang[b], Qiang Shen[b]

[a]*Information Technology College, Dalian Maritime University, Dalian, 116026, China*
[b]*Department of Computer Science, Institute of Mathematics, Physics and Computer Science, Aberystwyth University, SY23 3DB Aberystwyth, UK*

Abstract

Feature selection plays an important role in reducing irrelevant and redundant features, while retaining the underlying semantics of selected ones. An effective feature selection method is expected to result in a significantly reduced subset of the original features without sacrificing the quality of problem-solving (e.g., classification). In this paper, a non-unique decision measure is proposed that captures the degree of a given feature subset being relevant to different categories. This helps to represent the uncertainty information in the boundary region of a granular model, such as rough sets or fuzzy-rough sets in an efficient manner. Based on this measure, the paper further introduce a differentiation entropy as an evaluator of feature subsets to implement a novel feature selection algorithm. The resulting feature selection method is capable of dealing with either nominal or real-valued data. Experimental results on both benchmark data sets and a real application problem demonstrate that the features selected by the proposed approach outperform those attained by state-of-the-art feature selection techniques, in terms of both the size of feature reduction and the classification accuracy.

Keywords: Non-unique decision, Differentiation entropy, Feature selection.

*Corresponding author.
Email address: yanpengqu@dlmu.edu.cn (Yanpeng Qu)

1. Introduction

With the augment of the dimensionality of information, the prediction performance, the learning efficiency, the data visualisation and the information comprehensibility will be degraded due to the existence of the irrelevant, redundant and noisy features. In general, there are two approaches to combat such problems: dimensionality reduction (DR) and feature (subset) selection (FS) [1]. DR transforms the data in the high-dimensional space to a space of fewer dimensions. The data transformation may be linear, as in principal component analysis (PCA) [2] and linear discriminant analysis [3], but many nonlinear dimensionality reduction techniques also exist, such as Sammon mapping [4] and Laplacian eigenmaps [5]. FS processes data to select those features that are most informative of a given result, which maintain the meaning of the features from an original set by eliminating irrelevant and redundant features. Three strategies exist that may be utilised to implement either approach: the filter strategy [6], the wrapper strategy [7], and the embedded strategy [8]. Since both DR and FS may even help increase the quality of the reduced data sets [9], they are widely used in many areas such as text categorisation [10, 11], plant monitoring [12], and patient treatment [13, 14].

Rough set theory is an effective tool to deal with incomplete, uncertain information [15]. Generally the extension of rough sets consists of fuzzy-rough sets [16] (e.g. vaguely quantified rough sets [17], kernelised fuzzy-rough set [18]), probabilistic rough sets [19] (e.g., decision-theoretic rough sets [20], variable precision rough sets [21]) and the rough sets based on the tolerance relation (e.g., neighbourhood rough set [22], covering rough sets [23]). It is noteworthy that due to the variety of fuzzy membership representations, certain fuzzy-rough sets, such as kernelised fuzzy-rough sets, employ the tolerance relations also. One of the applications of rough sets and the associated variances is to identify feature (subset) dependency or uncertainty to the class decision. Either of these two measures has proven to be a decent indicator to implement FS [24].

Amongst different approaches to FS, the filter-based is by far the most popular when implemented with rough or fuzzy-rough methods. Following this FS strategy, in order to obtain a feature subset for a given problem, many searching mechanims have been employed. In [25], for example, a positive region-based FS algorithm is presented, where the significance of conditional features for the decision is measured by feature dependency between

them. Also, a rough entropy-based uncertainty measure is proposed in [26] to perform FS via evaluating the roughness and accuracy of the knowledge embedded in the data. However, these FS algorithms are typically focused on dependency degree in the positive region of a rough set, and the uncertainty information remains in the boundary region of a rough set is neglected. In [27], a boundary region-based algorithm is reported to evaluate the feature subsets, enabling the algorithm to find a reduct more effectively.

In this paper, a general non-unique decision measure (NDM) is presented to depict the inconsistency of a conditional feature subset for decision category distinguishing. In particular, for nominal data, this measure degenerates into counting the number of feature values that may lead to different decisions. For real-valued data, NDM takes the form of an aggregation of the degrees measuring the implication between the equivalence classes induced by the feature subsets regarding the decision. It implementation, when such an aggregator is set to be the *maximum* operator, NDM collapses to the amount of uncertainty information remaining in the fuzzy-rough boundary region. Based on NDM, a differentiation entropy (NDE) is proposed as a feature subset quality evaluator to implement FS. With the degradation of the NDE value, the significance of the corresponding feature subset becomes more representative for all the original features within the information system concerned.. This NDE-based FS method works because NDE exploits a more comprehensive understanding of the uncertainty information that contained within conventional rough set-based dependency measures.

In order to demonstrate the efficacy of the proposed FS method, comparative experimental studies are carried out on both benchmark data sets (for nominal data) and a real-world application, regarding mammographic risk assessment [28] (involving real-valued data). The work is compared with popular state-of-the-art FS and other DR techniques, including ACOFS [29], CFS [30], RFS [25], PCA [2] and CSFS [31]. It is shown that the proposed algorithm outperforms the rest, returning a high classification accuracy for the benchmark data sets investigated.

The remainder of this paper is structured as follows. In Section 2, the preliminary of rough set theory, fuzzy-rough set theory and differentiation entropy is reviewed. Section 3 introduces the concept of NDM and the associated NDE, and describes the NDE-based FS algorithm. In Section 4, the comparative experimental results are presented and discussed. Section 5 summarises the paper and points out interesting further work.

2. Background

This section reviews the mathematical concepts concerning rough sets, fuzzy-rough sets which are relevant to the FS process developed in this work, and introduces the basic ideas of differentiation entropy.

2.1. Rough set theory

Let (U, A) be an information system, where U is a non-empty finite set of objects, and A is a non-empty finite set of features such that $a : U \to V_a$ for every $a \in A$. Where V_a is the set of values that the feature a can take. The information system (U, A) can also be defined as a decision table by $(U, C \cup D)$, where $C \cup D = A, C \cap D = \varnothing$, and C is the set of conditional features and D is the set of decision attributes, respectively. For each feature subset $P \subseteq C$ an associated indistinguishable relation can be determined:

$$IND(P) = \{(x, y) \in U^2 \mid \forall a \in P, a(x) = a(y)\}. \tag{1}$$

where $a(x)$ is the value of an object x on the feature a.

Obviously, $IND(P)$ is an equivalence relation on U. The partition of U determined by $IND(P)$ is herein denoted by U/P which can be defined such that

$$U/P = \otimes\{U/a|a \in P\}. \tag{2}$$

where $\otimes$ is defined as follows for sets A and B:

$$A \otimes B = \{X \cap Y|X \in A, Y \in B, X \cap Y \neq \varnothing\}. \tag{3}$$

For any object $x \in U$, the equivalence class determined by $IND(P)$, is denoted by $[x]_P$. For any $X \subseteq U$ and $P \subseteq C$, the P-lower and P-upper approximations of X are respectively defined as:

$$\underline{P}X = \{x \in U \mid [x]_P \subseteq X\}. \tag{4}$$

$$\overline{P}X = \{x \in U \mid [x]_P \cap X \neq \varnothing\}. \tag{5}$$

Informally, the former depicts the set of those objects which can be said with certainty to belong to the concept to be approximated, and the latter idoes the set of objects which either definitely or possibly belong to the concept to be approximated. The difference between the upper and lower approximation is the area known as the boundary region and thus, represents the area of

uncertainty. When the boundary region is empty, there is no uncertainty regarding the concept which is being approximated and all objects belong to the subset of objects of interest with full certainty.

Thanks to the introduction of equivalence relations, the universe U can be partitioned into such different regions. Particularly, given the feature subsets P and Q with respect to C, the important concepts of positive, negative and boundary regions can be defined respectively as:

$$POS_P(Q) = \bigcup_{X \in U/Q} \underline{P}X, \tag{6}$$

$$NEG_P(Q) = U - \bigcup_{X \in U/Q} \overline{P}X, \tag{7}$$

$$BND_P(Q) = \bigcup_{X \in U/Q} \overline{P}X - \bigcup_{X \in U/Q} \underline{P}X. \tag{8}$$

Interestingly, in conventional rough set-based approach to FS [10], by employing the above concept of positive region, it is possible to calculate the degree of dependency of a feature set Q upon another P. In particular, for P, $Q \subseteq A$, it can be said that Q depends on P in a degree k $(0 \leq k \leq 1)$, which is defined as follow:

$$k = \gamma_P(Q) = \frac{\mid POS_P(Q) \mid}{\mid U \mid}. \tag{9}$$

2.2. Fuzzy-rough set theory

Fuzzy-rough sets are a fuzzy extension of rough sets [16]. In a fuzzy-rough set, the two types of approximation in rough sets are both fuzzified, leading to fuzzy lower and upper approximations. Definitions for the fuzzy lower and upper approximations can be found in [16, 25], where a T-transitive fuzzy similarity relation is used to approximate a fuzzy concept X:

$$\mu_{\underline{R_P}X}(x) = \inf_{y \in U} I(\mu_{R_P}(x, y), \mu_X(y)), \tag{10}$$

$$\mu_{\overline{R_P}X}(x) = \sup_{y \in U} T(\mu_{R_P}(x, y), \mu_X(y)). \tag{11}$$

In the above, U is a nonempty set of finite objects; I is a fuzzy implicator; T is a T-norm; R_P is the fuzzy similarity relation induced by the subset of

features P:

$$\mu_{R_P}(x,y) = T_{a\in P}\{\mu_{R_a}(x,y)\}. \tag{12}$$

$\mu_{R_a}(x,y)$ is the degree to which objects x and y are similar with respect to feature a, and may be defined in many ways, for example:

$$\mu_{R_a}(x,y) = 1 - \frac{|a(x)-a(y)|}{|a_{max}-a_{min}|}, \tag{13}$$

$$\mu_{R_a}(x,y) = \max\left(\min\left(\frac{(a(y)-(a(x)-\sigma_a))}{(a(x)-(a(x)-\sigma_a))}, \frac{((a(x)+\sigma_a)-a(y))}{((a(x)+\sigma_a)-a(x))}\right), 0\right), \tag{14}$$

where ${\sigma_a}^2$ is the statistical variance of the feature a.

Given these definitions regarding fuzzy-rough lower and upper approximations, the fuzzy-rough boundary region for a fuzzy concept X can be introduced, such that

$$\mu_{BND_{R_P}X}(x) = \mu_{\overline{R_P}X}(x) - \mu_{\underline{R_P}X}(x). \tag{15}$$

The uncertainty for a concept X using features in P can therefore be calculated as follows:

$$\mu_P(X) = \frac{\sum_{x\in\mathbb{U}} \mu_{BND_{R_P}X}(x)}{|U|}. \tag{16}$$

Indeed, the value of such indicator is the average extent to which objects belong to the fuzzy boundary region for the concept X.

The total uncertainty degree for all concepts, which are based on the equivalence relations over U induced by the decision attribute set Q upon a conditional feature subset P, is defined by

$$\lambda_P(Q) = \frac{\sum_{X\in U/Q} U_P(X)}{|U/Q|}. \tag{17}$$

2.3. Differentiation entropy

The notion of differentiation entropy is proposed to facilitate measuring the difference between the partition induced by a certain feature subset and that by all features [32]. Significantly, a number of important properties can be derived from this uncertainty measure. Formally, let $(U, C \cup D)$ be an information decision system as defined previously and $P \subseteq C$, the

differentiation entropy of P with respect to C is defined by

$$E(P|U \oplus C) = -\frac{1}{U}\sum_{x\in U} log_2 \frac{|[x]_C \cap [x]_P|}{|[x]_P|}. \tag{18}$$

This entropy measure represents the difference of the discernibility of the information between a feature subset and the full feature set. Thus, it provides a way to gauge the discernibility over the knowledge embedded in the original data.

Note that, for any $P \subseteq C$ and $x \in U$, there is $[x]_C \subseteq [x]_P$ and

$$E(P|U \oplus C) = -\frac{1}{U}\sum_{x\in U} log_2 \frac{|[x]_C|}{|[x]_P|}. \tag{19}$$

Therefore, for $P \subseteq B \subseteq C$, the following properties of the differentiation entropy hold:

$$E(B|U \oplus C) \leqslant E(P|U \oplus C), \tag{20}$$

$$E(B|U \oplus C) = E(P|U \oplus C), U/P = U/B. \tag{21}$$

Given these properties, it can be seen that $E(P|U \oplus C)$ not only shows the difference between U/P and U/C regarding their respective overall description ability, but also reflects the significance of the feature subset P with respect to the entire original feature set C. Indeed, the larger the value of $E(P|U \oplus C)$, the greater the difference of the feature subset P in representing C becomes. In particular, if $E(P|U \oplus C) = 0$, then $U/P = U/C$. This implies that the same knowledge description power between U/P and U/C results, which in turn, means that the significance of the reduced feature subset P is equivalent to that of the entire feature set C.

3. Non-unique decision-based differentiation entropy

This section introduces the concept of NDM and the associated NDE measure, based upon which the section also puts forward a novel FS algorithm.

3.1. NDE for nominal data

The concept of unique decision for nominal data is introduced in [33] in an effort to optimise the degree of dependency defined by exploiting the

concept of positive region in rough sets. However, since the degree of dependency can only provide information from the positive region, the information contained within the boundary region is neglected. Having recognised this, in this paper, the notion of an NDM is proposed to effectively characterise the uncertainty information resided in this region.

Let $(U, C \cup D)$ be an information decision system as specified earlier, for any $P,\ Q \subseteq C \cup D$. Define a non-deducible or inconsistent relation over U between Q and P as

$$\tau_P Q = |U/P - (U/P \otimes U/Q)|\,. \tag{22}$$

From this definition, if Q is the decision attribute set D, for conditional feature subset P, $\tau_P D$ represents the total number of feature values that lead to a non-unique decision using P. This therefore, captures the same information as with the boundary region of a rough set. With the introduction of τPD, the NDM for D on P can be rewritten as

$$NDM_P = \frac{\tau_P D}{|U|}. \tag{23}$$

It can be readily established that $0 \leqslant NDM_P \leqslant 0.5$. If $NDM_P = 0$, this means that the indistinguishable relation $IND(P)$ can be used to classify each sample into a distinct decision. If $NDM_P = 0.5$ it implies that each $[x]_P$ only contains two samples and is not subsumed by any $[x]_D$, and hence, that the number of $[x]_P$ is $|U|/2$. More generally, the value of NDM provides a measure over any inconsistency of a given conditional feature subset P for decision-making. In particular, if features subsets $P \subseteq B \subseteq C$, then $[x]_B \subseteq [x]_P$. Therefore, it can be observed that $\tau_B D \leqslant \tau_P D$ and $NDM_B \leqslant NDM_P$, given $P \subseteq B$.

From the above, for any $P \subseteq C$, the differentiation entropy of P with respect to C, and also to D (owing to the embedment of D in $\tau_P D$) can then be defined as:

$$E(P|U \oplus C) = -\frac{1}{|U|} \sum_{x \in U} log_2 \frac{NDM_C + 1}{NDM_P + 1}. \tag{24}$$

In this case, for any $P \subseteq B \subseteq C$, there is $E(B|U \oplus C) \leqslant E(P|U \oplus C)$. Thus, $E(P|U \oplus C)$ monotonically decreases while the number of the features in the subset increases. As $0 \leqslant E(P|U \oplus C)$, the optimal selection for the

feature subset P can be determined by minimising $E(P|U \oplus C)$. Therefore, following the use of standard rough set-based FS terminology [10], for any $P \subseteq C$, if $E(P|U \oplus C) = 0$ and for every $p \in P$, $E(P - \{p\}\,|U \oplus C) \neq 0)$, P is called a reduct of C with respect to D.

3.2. NDE for real-valued data

As introduced above, for nominal data, NDM reflects the total number of feature values leading to non-unique decisions. Generalising this definition to cope with real-valued data intuitively, NDM is herein designed to be an aggregation of the degrees to which those fuzzy equivalence classes induced by the decision D fail to be implicated by the fuzzy equivalence classes of a given sample x.

Formally, given an information decision system $(U, C \cup D)$, the inconsistency measure for the decision D on feature subset $P \subseteq C$ is defined by

$$\tau_P D(x) = \underset{y \in U}{A} \left\{1 - I\left(\mu_{R_P}(x,y), \mu_{R_D}(x,y)\right)\right\}, \tag{25}$$

where A is an aggregation operator ranging from 0 to 1; I is a fuzzy implicator; μ_{R_p} is the fuzzy similarity function defined in Eq.(12); $\mu_{R_D}(x,y)$ is 1 when x and y have the identical classification decision, or 0 otherwise. Note that when A is set to be the S-norm *maximum* (or *supremum*), $\tau_P D(x)$ collapses to the fuzzy-rough boundary region of the decision D (Eq. (15)) where the upper approximation (11) is equal to 1. Thus, $\tau_P D(x)$ reflects the uncertain information contained within boundary region.

From this, the non-unique decision measure using features in P can be calculated as follows:

$$NDM_P = \frac{\sum_{x \in U} \tau_P D(x)}{|U|}. \tag{26}$$

Thus, the differentiation entropy of P with respect to C and D can be defined as:

$$E(P|U \oplus C) = -\frac{1}{|U|} \sum_{x \in U} log_2 \frac{NDM_C + 1}{NDM_P + 1}. \tag{27}$$

Similar to the crisp version, for any $P \subseteq B \subseteq C$, as $\mu_{R_B}(x,y) \leqslant \mu_{R_P}(x,y)$, it is induced that $\tau_B D(x) \leqslant \tau_P D(x)$ and then $E(B|U \oplus C) \leqslant E(P|U \oplus C)$. Thus, $E(P|U \oplus C)$ is inversely proportional to the number of the features in the subset.

3.3. Feature selection using NDE

Based on the two definitions of NDE proposed above, a feature selection approach is derived here.

3.3.1. For nominal data

The value of NDM can be implemented in an incremental manner. In particular, for a feature subset $P \subseteq C$, given a new sample:

- If the new sample is identical to N existing objects but with a distinct decision, $NDM_P = \frac{\tau_P D+N}{|U|+1}$.
- If the new sample is distinct to any existing object with respect to P, $NDM_P = \frac{\tau_P D}{|U|+1}$.

Algorithm 1 summarises the above intuitions to calculate NDM for nominal features, starting from an initial object set U_0. For each iteration, the number of the objects within the boundary region is derived from the $\tau_P D$ computed in response to the addition of any new object. The time complexity of this algorithm is $O(|C| \times |U|)$.

Algorithm 1 NDM_P for nominal data

Input: $DT = (U, C \cup D)$, $U = U_0 \cup U'$, $P \subseteq C$,
Output: NDM_P

for each $x' \in |U'|$ **do**
  $\forall\ T \in U_0/P \cap (U_0/P \otimes U_0/D)$
  $U_0 = U_0 \cup \{x'\}$
  $\forall\ S \in (U_0/P - (U_0/P \otimes U_0/D))$.
  if $x' \in S$ and $\exists x_i \in T$, s.t. $x_i = x'$, $i = 1, \ldots, N$, **then**
    $NDM_P = \frac{\tau_P D+N}{|U|+1}$
  else if $x' \notin T$ **then**
    $NDM_P = \frac{\tau_P D}{|U|+1}$
  end if
end for

3.3.2. For real-value data

Given Eqs. (25) and (26), Algorithm 2 can be derived following a similar approach to Algorithm 1, in an effort to compute NDM with respect to a real-valued feature subset P and categoric decision D. The time complexity

of this algorithm is $O(|U|^2)$. In particular, if the aggregator A is set to be the S-norm *maximum*, the computation regarding an object y can be simplified in the range of $U - [x]_D$ dictated by another object x. In this case, the time complexity for real-value data can be reduced to $O(|U| \times (|U| - |[x]_D|)$. Meanwhile, compared to boundary region-based fuzzy-rough FS, the proposed method avoids computing the upper approximation. Thus, the overall efficiency of this algorithm can be improved considerably.

Algorithm 2 NDM_P for real-valued data

Input: $DT = (U, C \cup D)$, $P \subseteq C$,
Output: NDM_P

1: **for** each $x \in U$ **do**
2: **for** each $y \in U$ **do**
3: $\tau_P D(x) = \underset{y \in U}{A} \{1 - I(\mu_{R_P}(x, y), \mu_{R_D}(x, y))\}$
4: **end for**
5: $NDM_P = \frac{\sum_{x \in U} \tau_P D(x)}{|U|}$
6: **end for**

By computing NDM for each feature subset, the FS process based on the computation of NDE results, as shown in Algorithm 3. This method, shorthanded as NDEFS (standing for NDE-based FS) searches for the smallest feature subset whose NDE is equal to 0. Given the time consumed by calculating NDM, the time complexity of Algorithm 3 is $O(|C|^2 \times |U|)$.

Algorithm 3 NDE-based Feature Selection

Input: $DT = (U, C \cup D)$, $P \subseteq C$,
Output: R

1: Initialise $R = \varnothing$, $E_0 = 1$
2: **for** $\forall a \in C - R$ **do**
3: **if** $0 < E(R \cup \{a\} | U \oplus C) \leqslant E_0$ **then**
4: $E_0 = E(R \cup \{a\} | U \oplus C)$
5: **else**
6: Return R and Break
7: **end if**
8: $R = R \cup \{a\}$
9: **end for**

4. Experimental Evaluation

In this section, comparative experimental investigations are reported to evaluate the results of reduced data sets, in terms of the size of returned feature subsets and the running accuracy of employing them (when used to perform classification tasks). The experiments are run on eight nominal data sets taken from UCI repository of machine learning databases [34] and four real-valued data sets for mammographic risk assessment [28].

4.1. On nominal data sets

As indicated above, experimental results are herein discussed from two aspects: feature subset reduction and classification effectiveness.

4.1.1. Comparison on reduct size

Table 1 summarises the data sets used to conduct this experiment. Table 2 presents the experimental results in terms of reduced data set size gained by NDEFS are compared to those of state-of-the-art FS methods, such as ACOFS [29], CFS [30], RFS [25] and PCA [2]. These results show that NDEFS outperforms other popular FS methods on most of the eight data sets. For example, on the data set *spectf*, NDEFS selects only 2 features as a reduct, while the reducts returned by the alternatives are much larger than it. Considering the average size of the reducts returned by all FS methods, NDEFS results in the best performance as well. Note that whilst ACOFS, CFS and RFS automatically determine the number of selected features, the size of each returned subset by PCA is empirically determined with respect to the best classification accuracy achievable as a certain number of principal components is taken. The following further investigation into the accuracy of using selected feature subsets will show that the returned reducts by NDEFS also retain sufficient information to entail high discriminating ability.

4.1.2. Comparison on classification accuracy

The classification accuracies achievable using the reduced data sets are compared here, again amongst NDEFS, ACOFS, CFS, FRFS [25] and PCA. For completeness, the classification methods used in this paper are briefly summarised as follows.

- NB (Naive Bayes) [35] is a simple probabilistic classifier, directly applying Bayes theorem [36] with strong (naive) independence assumptions. Depending on the precise nature of the probability model used, naive

Table 1: Eight nominal benchmark data sets

Data set	Objects	Features
coil	9822	86
credit	1000	25
handwritten	1953	257
satellite	6435	37
spect	187	23
spectf	187	45
wisconsin	683	10
zoo	101	17

Table 2: Reduct sizes for benchmark data sets

Data set	NDEFS	ACOFS	CFS	RFS	PCA
coil	28	73	9	33	46
credit	5	17	6	13	20
handwritten	22	21	74	22	162
satellite	7	19	24	15	6
spect	12	15	9	15	18
spectf	2	6	9	6	24
wisconsin	4	7	9	7	7
zoo	6	8	9	7	10
average	10.75	20.75	18.63	14.75	36.63

Bayesian classifiers can be trained very efficiently in a supervised learning setting. The training only requires a small amount of training data to estimate the parameters (means and variances of the variables) necessary for classification.

- SMO (sequential minimal optimisation) [37] is an algorithm for efficiently solving optimisation problems which arise during the training of a support vector machine [38]. It breaks optimisation problems into a series of smallest possible sub-problems, which are then resolved analytically.

- J48 (decision tree based classifier) [39] creates decision trees by choosing the most informative features and recursively partitioning a training data table into subtables based on the values of such features. Each node

in the tree represents a feature, with the subsequent nodes branching from the possible values of this node according to the current subtable. Partitioning stops when all data items in the subtable have the same classification. A leaf node is then created to represent this classification. Incidentally, J48 is a representative of the wrapped approach to feature selection as it is capable of generalising the training data into a decision tree that involves only a subset of the original features.

To demonstrate the validity of the feature subsets selected by NDEFS, Tables 3, 4 and 5 show the classification accuracy rates produced by the above three classification algorithms, respectively. The last column of each table lists the results for unreduced original data sets. 10×10-fold cross-validation is constructed throughout the experimentation. The advantage of running cross validation over random sub-sampling is that all objects are used for both training and testing, and each object is used for testing only once per fold [40].

Table 3: Classification accuracy by NB

Data set	NB					
	NDEFS	ACOFS	CFS	RFS	PCA	Unred
coil	89.77	76.59	92.60	86.86	77.33	77.98
credit	71.18	76.29	75.24	75.46	71.82	75.59
handwritten	69.45	59.00	85.76	69.45	48.93	86.19
satellite	78.23	79.33	79.54	78.49	77.69	79.59
spect	86.89	86.25	88.21	80.95	82.26	80.92
spectf	90.51	74.45	69.30	78.46	64.69	66.89
wisconsin	96.47	96.41	96.34	96.73	96.08	96.34
zoo	94.84	97.83	95.95	94.90	89.55	96.95

Take Table 3 as an example. It can be seen from the table that in conjunction with the use of NB, NDEFS leads to better classification accuracy results for several data sets. In particular, the accuracy rates over the data sets *coil* and *spectf* generated by NDEFS are superior to those generated by all other FS methods. This is achieved through the use of the smallest feature subsets returned by it. For those data sets where the use of NDEFS-returned features do not lead to the highest accuracy, the performances remain compatible to the rest, but mostly involving far less features.

Table 4: Classification accuracy by SMO

Data set	SMO					
	NDEFS	ACOFS	CFS	RFS	PCA	Unred
coil	94.03	94.03	94.03	94.03	94.03	94.03
credit	70.00	76.94	75.15	75.92	73.26	76.72
handwritten	74.98	62.78	89.10	74.98	53.57	93.58
satellite	84.19	85.49	85.86	85.46	82.27	86.78
spect	92.02	92.02	92.02	92.02	92.02	91.96
spectf	92.02	92.02	92.02	92.02	92.02	92.02
wisconsin	96.72	96.62	97.01	97.07	96.59	97.01
zoo	82.39	90.84	94.49	92.49	89.65	93.68

Table 5: Classification accuracy by J48

Data set	J48					
	NDEFS	ACOFS	CFS	RFS	PCA	Unred
coil	93.96	93.93	94.03	93.97	94.01	93.92
credit	68.91	72.81	73.05	73.13	69.26	73.57
handwritten	67.11	58.84	76.17	67.11	51.29	76.13
satellite	85.24	86.71	86.50	86.17	83.87	86.41
spect	92.02	92.02	92.02	92.02	92.02	92.02
spectf	91.65	91.91	88.16	90.90	86.11	84.99
wisconsin	96.91	95.85	95.44	95.36	95.43	95.44
zoo	93.83	94.03	93.08	95.15	90.45	92.61

Tables 4 and 5 show similar observations to Table 3. For instance, when J48 is employed for classification (see Table 5), the results using NDEFS over the data sets *wisconsin* and *monk*3 are consistently better than those of the alternatives. Again, this performance is obtained with the use of less features. Occasionally, NDEFS does not lead to a top classification rate. However, for such cases, it does not lead to the poorest performance either, producing generally well above average accuracy across the compared methods. Together, these results illustrate that the proposed approach has a better overall performance in terms of both classification accuracy and feature subset size.

4.2. Mammographic risk assessment

The data employed in this experimental evaluation is derived from the mammographic image analysis society (MIAS) database [41] (see [28] for the feature extraction process). It includes a complete data set of Medio-Lateral-Oblique (MLO) left and right mammograms of 161 women (322 objects). Each mammgram object is reprented by 280 features, in which 10 derived from morphological characteristics, and the remaining 270 derived from the extracted texture information. The spatial resolution of the images is $50\mu m \times 50\mu m$, quantised to 8 bits, and the linear optical density range is 0-3.2. Mammography images commonly used to perform risk assessment are based on the BI-RADS [42], Boyd [43], Tabar [44] or Wolfe [45] labelling schemes.

Table 6 shows the reduced feature subset size of the mammographic data sets using NDEFS, with respect to the aforementioned four labelling strategies, respectively. These reduced data sets are used in the comparative study below.

Table 6: Reduct size of MIAS data sets

Data set	NDEFS	ACOFS	CFS	CSFS	PCA
BI-RADS	7	7	35	15	12
Boyd	6	8	32	14	12
Tabár	6	7	31	15	12
Wolfe	6	7	30	14	12
average	6.25	7.25	32	14.5	12

As with the experiments on nominal-valued benchmark datasets, stratified 10×10-fold cross-validation is also used herein for all the four different labelling strategies. Also, comparisons are made again amongst the use of Unred (i.e., the unreduced original datasets) and that of those returned by NDEFS, CFS, CSFS [31], PCA and ACOFS, via running the same classifiers described previously.

The classification accuracy rates of the reduced MIAS data sets are reported in Tables 7, 8 and 9, respectively for the three classifiers. Generally, the data sets reduced by NDEFS give the best results. Especially, for the reduced Boyd data set, the classification task conducted with NDEFS-returned feature subset results in the best performance. Together with the previous results, overall, it is clear that NDEFS can effectively select less features while leading to a better classification performance.

Table 7: Classification accuracy by NB

Data set	NB					
	NDEFS	ACOFS	CFS	CSFS	PCA	Unred
BI-RADS	72.05	68.94	72.36	70.19	66.15	70.81
Boyd	53.42	56.52	59.01	57.45	50.93	57.45
Tabár	59.01	59.63	61.80	59.94	59.94	58.70
Wolfe	62.42	65.21	68.01	66.15	59.63	65.53

Table 8: Classification accuracy by SMO

Data set	SMO					
	NDEFS	ACOFS	CFS	CSFS	PCA	Unred
BI-RADS	71.74	71.43	75.77	73.29	69.25	73.60
Boyd	57.45	56.83	58.38	61.49	58.69	59.32
Tabár	59.01	59.32	68.01	63.35	64.90	66.46
Wolfe	65.53	66.15	69.87	69.57	65.22	70.50

Table 9: Classification accuracy by J48

Data set	J48					
	NDEFS	ACOFS	CFS	CSFS	PCA	Unred
BI-RADS	63.66	68.63	71.43	66.46	67.39	65.84
Boyd	50.00	48.76	51.24	53.10	50.00	48.75
Tabár	53.73	51.55	59.63	59.01	59.31	59.32
Wolfe	56.21	59.94	63.97	63.35	59.62	62.42

5. Conclusion

This paper has presented a non-unique decision measure to evaluate the uncertainty of a feature subset for use in support of classification. Particularly, the work utilises differentiation entropy to examine the difference between an emerging feature subset and the original full set of features, identifying an optimal feature subset that contains sufficient information for maintaining the discriminating ability of the original features. The proposed FS algorithm has been fully implemented and tested against popular, state-of-the-art FS methods on both nominal-valued benchmark data sets and real-valued data sets, with the latter in the context of addressing real-world problems of mammographic risk assessment. Comparative experimental results have demonstrated in general that the proposed FS approach can identify fea-

ture subsets of much smaller in size than those competing existing methods, and that the proposed FS algorithm can lead to the achievement of higher classification accuracy.

Topics for further research include a more comprehensive development of the FS method to handle more complicated large-scale data sets [46], including mixed forms of both nominal and real-valued data. In addition, how this work may be extended to deal with non-boolean classification tasks is also very interesting. Last but not least, potential alternative applications of the proposed NDM in unsupervised feature selection [47], fuzzy-rough classification [48, 49], classification ensembles [50], parallel computing [51, 52] or uncertain data query [53, 54], remain active research.

Acknowledgments

This work is jointly supported by the National Natural Science Foundation of China (No. 61502068), the China Postdoctoral Science Foundation (No. 2013M541213 and 2015T80239), the Royal Society International Exchanges Cost Share Award with NSFC (No. IE160875), and a Sêr Cymru II COFUND Fellowship, UK. The authors would like to thank the anonymous referees for their constructive comments which have been very helpful in revising this paper.

References

[1] C. A. Murthy, Bridging feature selection and extraction: Compound feature generation, IEEE Transactions on Knowledge and Data Engineering 29 (4) (2017) 757–770.

[2] R. Bro, A. K. Smilde, Principal component analysis, Analytical Methods 6 (9) (2014) 2812-2831.

[3] A. J. Izenman, Linear Discriminant Analysis, Springer New York, New York, NY, 2008, pp. 237–280.

[4] J. W. Sammon, A nonlinear mapping for data structure analysis, IEEE Transactions on Computers C-18 (5) (1969) 401–409.

[5] M. Belkin, P. Niyogi, Laplacian eigenmaps for dimensionality reduction and data representation, Neural Computation 15 (6) (2003) 1373–1396.

[6] J. Xu, B. Tang, H. He, H. Man, Semisupervised feature selection based on relevance and redundancy criteria, IEEE Transactions on Neural Networks and Learning Systems 28 (9) (2017) 1974–1984.

[7] G. Chen, J. Chen, A novel wrapper method for feature selection and its applications, Neurocomputing 159 (2015) 219–226.

[8] J. Zhao, L. Chen, W. Pedrycz, W. Wang, Variational inference-based automatic relevance determination kernel for embedded feature selection of noisy industrial data IEEE Transactions on Industrial Electronics 66 (1) (2019) 416–428.

[9] R. Jensen, Q. Shen, Are more features better? IEEE Transactions on Fuzzy Systems 17 (6) (2009) 1456–1458.

[10] A. Chouchoulas and Q. Shen. Rough set-aided keyword reduction for text categorisation. Applied Artificial Intelligence, 15 (9):(2001) 843–873.

[11] H. Uğuz, A two-stage feature selection method for text categorization by using information gain, principal component analysis and genetic algorithm, Knowledge-Based Systems 24 (7) (2011) 1024–1032.

[12] Q. Shen, R. Jensen, Selecting informative features with fuzzy-rough sets and its application for complex systems monitoring. Pattern Recognition, 37 (7) (2004) 1351–1363.

[13] J. Chen, K. Li, Z. Tang, K. Bilal, K. Li, A parallel patient treatment time prediction algorithm and its applications in hospital queuing-recommendation in a big data environment, IEEE Access 4 (2016) 1767–1783.

[14] J. Chen, K. Li, Z. Tang, K. Bilal, S. Yu, C. Weng, K. Li, A parallel random forest algorithm for big data in a spark cloud computing environment, IEEE Transactions on Parallel and Distributed Systems 28 (4) (2017) 919–933.

[15] Z. Pawlak, Rough sets, International Journal of Computer & Information Sciences 11(5) (1982) 341–356.

[16] R. Jensen, Q. Shen, Computational intelligence and feature selection: rough and fuzzy approaches, Vol. 8, John Wiley & Sons, 2008.

[17] C. Cornelis, M. De Cock, A. Radzikowska, Vaguely quantified rough sets, Lecture Notes in Artificial Intelligence 4482 (2007) 87–94.

[18] Y. Qu, C. Shang, Q. Shen, N. Mac Parthaláin, W. Wu, Kernel-based fuzzy-rough nearest-neighbour classification for mammographic risk analysis, International Journal of Fuzzy Systems 17 (3) (2015) 471–483.

[19] Y. Yao, S. Greco, R. Słowiński, Probabilistic Rough Sets, Springer Berlin Heidelberg, Berlin, Heidelberg, 2015, pp. 387–411.

[20] G. Lang, D. Miao, M. Cai, Three-way decision approaches to conflict analysis using decision-theoretic rough set theory, Information Sciences 37–407 (2017) 185–207.

[21] Y. Yao, J. Mi, Z. Li, A novel variable precision (θ, σ)-fuzzy rough set model based on fuzzy granules, Fuzzy Sets and Systems 236 (2014) 58–72.

[22] Y. Chen, Y. Xue, Y. Ma, F. Xu, Measures of uncertainty for neighborhood rough sets. Knowledge-Based Systems 120 (2017) 226-235.

[23] L. Ma, The investigation of covering rough sets by Boolean matrices, International Journal of Approximate Reasoning 100 (2018) 69–84.

[24] R. Jensen, Q. Shen, Semantics-preserving dimensionality reduction: Rough and fuzzy-rough approaches. IEEE Transactions on Knowledge and Data Engineering, 16 (12) (2004) 1457–1471.

[25] R. Jensen, Q. Shen, New approaches to fuzzy-rough feature selection, IEEE Transactions on Fuzzy Systems 17 (4) (2009) 824–838.

[26] L. Sun, J. Xu, Y. Tian, Feature selection using rough entropy-based uncertainty measures in incomplete decision systems, Knowledge-Based Systems 36 (2012) 206–216.

[27] Z. Lu, Z. Qin, Y. Zhang, J. Fang, A fast feature selection approach based on rough set boundary regions, Pattern Recognition Letters 36 (2014) 81–88.

[28] A. Oliver, J. Freixenet, R. Marti, J. Pont, E. Perez, E. Denton, R. Zwiggelaar, A novel breast tissue density classification methodology, IEEE Transactions on Information Technology in Biomedicine 12 (1) (2008) 55–65.

[29] R. Jensen, Q. Shen, Fuzzy-rough data reduction with ant colony optimization, Fuzzy Sets and Systems 149 (1) (2005) 5–20.

[30] M. A. Hall, Correlation-based feature selection for machine learning, Ph.D. thesis, The University of Waikato (1999).

[31] H. Liu, R. Setiono, A probabilistic approach to feature selection – a filter solution, in: International Conference on Machine Learning, 1996, pp. 319–327.

[32] F. Li, Z. Zhang, C. Jin, Feature selection with partition differentiation entropy for large-scale data sets, Information Sciences 329 (2016) 690–700.

[33] M. S. Raza, U. Qamar, An incremental dependency calculation technique for feature selection using rough sets, Information Sciences 343 (2016) 41–65.

[34] C. Blake, C. Merz, UCI repository of machine learning databases, university of California, Irvine, School of Information and Computer Sciences (1998).

[35] C. R. Stephens, H. F. Huerta, A. R. Linares, When is the Naive Bayes approximation not so naive?, Machine Learning 107 (2) (2018) 397–441.

[36] Bayes' theorem in statistics, in: A. Papoulis (Ed.), Probability, Random Variables, and Stochastic Processes, 2nd Edition, 1984.

[37] X. Huang, L. Shi, J. Suykens, Sequential minimal optimization for SVM with pinball loss, Neurocomputing 149 (C) (2015) 1596–1603.

[38] C. Cortes, V. Vapnik, Support-vector networks, Machine Learning 20 (1995) 273–297.

[39] J. Quinlan, C4.5: Programs for Machine Learning, The Morgan Kaufmann Series in Machine Learning, Morgan Kaufmann, San Mateo, 1993.

[40] Y. Bengio, Y. Grandvalet, Bias in estimating the variance of K-fold cross-validation, in: Statistical Modeling and Analysis for Complex Data Problems, Springer, 2005, pp. 75–95.

[41] J. Suckling, J. Partner, D. Dance, S. Astley, I.Hutt, C. Boggis, I. Ricketts, E. Stamatakis, N. Cerneaz, S. Kok, D.Betal, P.Taylor, J. Savage, The mammographic image analysis society digital mammogram database, in: International Workshop on Digital Mammography, 1994, pp. 211–221.

[42] American College of Radiology, Illustrated Breast Imaging Reporting and Data System BIRADS, 3rd Edition (1998).

[43] N. Boyd, J. Byng, R. Jong, E. Fishell, L. Little, A. Miller, G. Lockwood, D. Tritchler, M. Yaffe, Quantitative classification of mammographic densities and breast cancer risk: results from the canadian national breast screening study, Journal of The National Cancer Institute 87 (9) (1995) 670–675.

[44] L. Tabár, T. Tot, P. Dean, The Art and Science of Early Detection with Mammography, Georg Thieme Verlag, 2005.

[45] J. Wolfe, Risk for breast cancer development determined by mammographic parenchymal pattern, Cancer 37 (1976) 2486–2492.

[46] C. Shang, Q. Shen, Aiding classification of gene expression data with feature selection: A comparative study, International Journal of Computational Intelligence Research 1 (1) (2001) 68–76.

[47] T. Boongoen, C. Shang, N. Iam-On, Q. Shen, Extending data reliability measure to a filter approach for soft subspace clustering. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 41 (6) (2011), 1705–1714.

[48] Y. Qu, Q. Shen, N. Mac Parthaláin, C. Shang, W. Wu, Fuzzy similarity-based nearest-neighbour classification as alternatives to their fuzzy-rough parallels, International Journal of Approximate Reasoning 54 (1) (2013) 184–195.

[49] Y. Qu, C. Shang, N. Mac Parthaláin, W. Wu, Q. Shen, Multi-functional nearest-neighbour classification, Soft Computing 22 (8) (2018) 2717–2730.

[50] R. Diao, F. Chao, T. Peng, N. Snooke, Q. Shen, Feature selection inspired classifier ensemble reduction, IEEE Transactions on Cybernetics, 44 (8) (2014) 1259–1268.

[51] K. Li, C. Liu, K. Li, A. Y. Zomaya, A framework of price bidding configurations for resource usage in cloud computing. IEEE Transactions on Parallel and Distributed Systems, 27 (8) (2016) 2168–2181.

[52] C. Liu, K. Li, C. Xu, K. Li, Strategy configurations of multiple users competition for cloud service reservation. IEEE Transactions on Parallel and Distributed Systems, 27 (2) (2016) 508–520.

[53] G. Xiao, K. Li, K. Li, X. Zhou, Efficient top-(k,l) range query processing for uncertain data based on multicore architectures. Distributed and Parallel Databases, 33 (3) (2015) 381–413.

[54] X. Zhou, K. Li, Y. Zhou, K. Li, Adaptive processing for distributed skyline queries over uncertain data, IEEE Transactions on Knowledge and Data Engineering, 28 (2) (2016) 371–384.