Expert Systems with Applications 38 (2011) 10737–10750

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Measuring relevance between discrete and continuous features based on neighborhood mutual information

Qinghua Hu [a,b], Lei Zhang [b,*], David Zhang [b], Wei Pan [b], Shuang An [a], Witold Pedrycz [c]

[a] *Harbin Institute of Technology, PO 458, Harbin 150001, PR China*
[b] *Department of Computing, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong, China*
[c] *Department of Electrical and Computer Engineering, University of Alberta, Canada*

ARTICLE INFO

Keywords:
Feature selection
Continuous feature
Relevance
Neighborhood entropy
Neighborhood mutual information

ABSTRACT

Measures of relevance between features play an important role in classification and regression analysis. Mutual information has been proved an effective measure for decision tree construction and feature selection. However, there is a limitation in computing relevance between numerical features with mutual information due to problems of estimating probability density functions in high-dimensional spaces. In this work, we generalize Shannon's information entropy to neighborhood information entropy and propose a measure of neighborhood mutual information. It is shown that the new measure is a natural extension of classical mutual information which reduces to the classical one if features are discrete; thus the new measure can also be used to compute the relevance between discrete variables. In addition, the new measure introduces a parameter delta to control the granularity in analyzing data. With numeric experiments, we show that neighborhood mutual information produces the nearly same outputs as mutual information. However, unlike mutual information, no discretization is required in computing relevance when used the proposed algorithm. We combine the proposed measure with four classes of evaluating strategies used for feature selection. Finally, the proposed algorithms are tested on several benchmark data sets. The results show that neighborhood mutual information based algorithms yield better performance than some classical ones.

1. Introduction

Evaluating relevance between features (attributes, variables) is an important task in pattern recognition and machine learning. In decision tree construction, indexes such as Gini, towing, deviance and mutual information were introduced to compute the relevance between inputs and output, thus guilding the algorithms to select an informative feature to split samples (Breiman, 1993; Quinlan, 1986, 1993). In filter based feature selection techniques, a number of relevance indexes were introduced to compute the goodness of features for predicting decisions (Guyon & Elisseeff, 2003; Hall, 2000; Liu & Yu, 2005). In discretization, a relevance index can be used to evaluate the effectiveness of a set of cuts by computing the effectiveness of a set of cuts by computing the relevance between the discretized features and decision (Fayyad & Irani, 1992, 1993; Liu, Hussain, & Dash, 2002). Relevance is also widely used in dependency analysis, feature weighting and distance learning (Düntsch & Gediga, 1997; Wettschereck, Aha, & Mohri, 1997).

In the last decades, a great number of indexes have been introduced or developed for computing relevance between features. Pearson's correlation coefficient, which reflects the linear correlation degrees of two random numerical variables, was introduced in Hall (2000). Obviously, there is some limitation in using this coefficient. First, correlation coefficient can just reflect the linear dependency between variables, while relations between variables are usually nonlinear in practice. Second, correlation coefficient cannot measure the relevance between a set of variables and another variable. In feature selection, we are usually confronted the task to compute the relation between a candidate feature and a subset of selected features. Furthermore, this coefficient may be not effective in computing the dependency between discrete variables. In order to address these problems, a number of new measures were introduced, such as mutual information (Battiti, 1994), dependency (Hu & Cercone, 1995; Pawlak & Rauszer, 1985) and fuzzy dependency in the rough set theory (Hu, Xie, & Yu, 2007), consistency in feature subset selection (Dash & Liu, 2003), Chi2 for feature selection and discretization (Liu & Setiono, 1997), Relief and ReliefF to estimate attributes (Sikonja & Kononenko, 2003). Dependency is the ratio of consistent samples which have the same decision if their values of inputs are the same over the whole set of training data. Fuzzy dependency generalizes this definition to the fuzzy condition. Consistency, proposed by Dash and Liu (2003), can be viewed as the ration of samples which can be correctly classified according to the majority decision.

[*] Corresponding author.
E-mail address: cslzhang@comp.polyu.edu.hk (L. Zhang).

0957-4174/$ - see front matter © 2011 Published by Elsevier Ltd.
doi:10.1016/j.eswa.2011.01.023

Among these measures, mutual information (MI) is the most widely used one in computing relevance. In ID3 and C4.5, MI is used to find good features for splitting samples (Quinlan, 1986, 1993). In feature selection, MI is employed to measure the quality of candidate features (Battiti, 1994; Fleuret, 2004; Hall, 1999; Hu, Yu, Xie, & Liu, 2006, Hu, Yu, & Xie, 2006; Huang, Cai, & Xu, 2008; Kwak & Choi, 2002, 2002; Liu, Krishnan, & Mondry, 2005; Peng, Long, & Ding, 2005; Qu, Hariri, & Yousif, 2005; Wang, Bell, & Murtagh, 1999; Yu & Liu, 2004). Given two random variables A and B, the MI is defined as

$$MI(A,B) = \sum_{a \in A} \sum_{b \in B} p(a,b) \log \frac{p(a,b)}{p(a)p(b)}.$$

Thus, MI can be considered as a statistics which reflects the degree of linear or nonlinear dependency between A and B. Generally speaking, one may desire that the selected features are highly dependent on the decision variable, but are independent between them. This condition makes the selected features maximally relevant and minimally redundant.

In order to compute mutual information, we should know the probability distributions of variables and their joint distribution. However, these distributions are not known in practice. Given a set of samples, we have to estimate the probability distributions and joint distributions of features. If features are discrete, histogram can be used to estimate the probabilities. The probabilities are computed as the relative frequency of samples with the corresponding feature values. If there are continuous variables, two techniques were developed. One is to estimate probabilities based on the technique of Parzen Window (Kwak & Choi, 2002; Wang et al., 1999). The other is to partition the domains of variables into several subsets with a discretization algorithm. From the theoretical perspective, the first solution is feasible. Whereas, it is usually difficult to obtain accurate estimates for multivariate density as samples in high-dimensional space is sparsely distributed. The computational cost is also very high (Liu et al., 2005; Peng et al., 2005). Considering the limit of Parzen Window, techniques of discretization are usually integrated with mutual information in feature selection and decision tree construction (C4.5 implicitly discretizes numerical variables into multiple intervals) (Hall, 1999; Liu et al., 2002; Qu et al., 2005; Yu & Liu, 2004). Discretization, as an enabling technique for inductive learning, is useful for rule extraction and concept learning (Liu et al., 2002). However, it is superfluous for C4.5, neural network and SVM. Moreover, discretization is not applicable to regression analysis, where relevance between continuous variables is desirable. In these cases an information measure for computing relevance between continuous features become useful.

In Hu, Yu, Liu, and Wu (2008), the authors considered that in human reasoning the assumptions of classification consistency are different in discrete and continuous feature spaces. In discrete spaces, the objects with the same feature values should be assigned with the same decision class; otherwise, we think the decision is not consistent. In the meanwhile, since the probability of two samples with the completely same feature values is very small in continuous spaces, we think the objects with the most similar feature values should belong to a decision class; otherwise, the decision is not consistent. The assumption of similarity in continuous spaces extends the one of equivalence in discrete spaces. Based on this assumption, Hu and his coworkers extended equivalence relation based dependency function to neighborhood relation based one, where neighborhood, computed with distance, is looked as the subset of samples which have the similar feature values with the centroid. Then by checking the purity of the neighborhood, we can determine whether the centroid sample is consistent or not. However, neighborhood dependency just reflects whether the sample is consistent, it is not able to record the degree of consistency of this sample; this makes the measure not so effective as mutual information in terms of stability and robustness. In this paper, we integrate the concept of neighborhood into Shannon's information theory, and propose a new information measure, called neighborhood entropy. Then, we derive the concepts of joint neighborhood entropy, neighborhood conditional entropy and neighborhood mutual information for computing the relevance between continuous variables and discrete decision features. Given this generalization, mutual information can be directly used to evaluate and select continuous features.

Our study is focused on three problems. First, we introduce the new definitions on neighborhood entropy and neighborhood mutual information. The properties of these measures are discussed. We show that the neighborhood entropy is a natural generalization of Shannon's entropy. Neighborhood entropy converts to the Shannon's one if a discrete distance is used.

Second, we discuss the problem how to use the proposed measures in feature selection. We give an axiomatic approach to feature subset selection and discuss the difference between the proposed one and other two approaches. In addition, we consider the ideas of maximal dependency, maximal relevance and minimal redundancy in the context of neighborhood entropy, and discuss their computational complexities. Finally, three strategies are proposed for selecting features based on neighborhood mutual information: maximal dependency (MD), minimal redundancy and maximal relevance (mRMR), minimal redundancy and maximal dependency (mRMD).

Finally, with comprehensive experiments, we exhibit the properties of neighborhood entropy and compare MD, mRMR and mRMD with some existing algorithms, such as CFS, consistency based feature selection, FCBF and neighborhood rough set based algorithm. The experimental results show the proposed measures are effective when being integrated with mRMR and mRMD.

The rest of the paper is organized as follows. Section 2 presents the preliminaries on Shannon's entropy and neighborhood rough sets. Section 3 introduces the definitions of neighborhood entropy and neighborhood mutual information and discusses their properties and interpretation. Section 4 integrates neighborhood mutual information with feature selection, where the relationships between MD, mRMR and mRMD are studied. Experimental analysis is described in Section 5. Finally, conclusion and future work are given in Section 6.

2. Preliminaries

2.1. Entropy and mutual information

Shannon's entropy, first introduced in 1948 (Shannon, 1948), is a measure of uncertainty of random variables. Let $A = \{a_1, a_2, \ldots, a_n\}$ be a random variable. If $p(a_i)$ is the probability of a_i, the entropy of A is defined as

$$H(A) = -\sum_{i=1}^{n} p(a_i) \log p(a_i).$$

If A and $B = \{b_1, b_2, \ldots, b_m\}$ are two random variables, the joint probability is $p(a_i, b_j)$, where $i = 1, \ldots, n$, $j = 1, \ldots, m$. The joint entropy of A and B is

$$H(A,B) = -\sum_{i=1}^{n} \sum_{j=1}^{m} p(a_i, b_j) \log p(a_i, b_j).$$

Assuming that the variable B is known, the uncertainty of A, named conditional entropy, is computed by

$$H(A|B) = H(A,B) - H(B) = -\sum_{i=1}^{n} \sum_{j=1}^{m} p(a_i, b_j) \log p(a_i, |b_j).$$

Correspondingly, the reduction of uncertainty of A resulting from the knowledge of B, called mutual information between A and B, is defined as

$$MI(A;B) = \sum_{i=1}^{n}\sum_{j=1}^{m} p(a_i, b_j)\log\frac{p(a_i|b_j)}{p(a_i)}.$$

As

$$\frac{p(a_i|b_j)}{p(a_i)} = \frac{p(b_j|a_i)}{p(b_j)} = \frac{p(a_i, b_i)}{p(a_i)p(b_j)},$$

so we have

$$\begin{aligned} MI(A;B) = MI(B;A) &= H(A) - H(A|B) = H(B) - H(B|A) \\ &= H(A) + H(B) - H(A,B). \end{aligned}$$

As to continuous random variables, the entropy is computed as

$$H(A) = -\int p(a)\log p(a)da,$$

where $p(a)$ is the probability density function.

In data-driven learning, the probability distributions of variables are usually unknown a priori. We have to estimate them making use of available samples.

2.2. Neighborhood rough sets

Given discrete data, the samples with the same feature value are pooled into a set, called equivalence class. These samples are expected to belong to the same class; otherwise, they are inconsistent. It is easy to verify whether the decisions are consistent or not by analyzing their decisions (Pawlak, 1991). However, it is unfeasible to compute equivalence classes with continuous features because the probability of samples with the same numerical value is very small. Intuitively speaking, the samples with the similar feature values should be classified into a single class in this case; otherwise, the decision is not consistent. Based on this observation, the model of neighborhood rough sets was proposed (Huang et al., 2008).

Given a set of samples $U = \{x_1, x_2, \ldots, x_n\}$, $x_i \in \Re^N$, Δ is a distance function on U, which satisfies $\Delta(x_i, x_j) \geqslant 0$; 2-norm distance (also called Euclidean distance): $\left(\sum_{k=1}^{N}|x_{ik} - x_{jk}|^2\right)^{1/2}$ is usually used in applications. Given $\delta \geqslant 0$, by $\delta(x) = \{x_i|\Delta(x, x_i) \leqslant \delta\}$, we denote the neighborhood of sample x_i. Given two feature spaces R and S, $\delta_R(x)$ and $\delta_S(x)$ are the neighborhoods of X computed in these feature spaces with infinite norm based distance, respectively. We have the following property: $\delta_{R\cup S}(x) = \delta_R(x) \cap \delta_S(x)$. In addition to the distance function given above, there are a number of distances for heterogeneous features and missing data (Wang, 2006).

Regarding a classification task, a decision attribute is given to assign a class label to each sample, and the samples are divided into $c_1, c_2, \ldots, c_k$, where $c_1 \cup c_2 \cdots \cup c_k = U$, and $c_i \cap c_j = \emptyset$ if $i \neq j$.

We say the decision of sample x is δ-neighborhood consistent if $\delta(x) \subseteq c_x$, where c_x is the subset of samples having the same class label as X. The consistent samples with class c_x is the lower approximation of c_x. Formally, the lower and upper approximations of decision class c_i are defined as

$$\underline{N}c_i = \{x|\delta(x) \subseteq c_i\}, \quad \overline{N}c_i = \{x|\delta(x) \cap c_i \neq \emptyset\},$$

respectively, where N denotes a neighborhood relation over U. Correspondingly, the total lower and upper approximations of classification C are written as

$$\underline{N}C = \bigcup_{i=1}^{k}\underline{N}c_i, \quad \overline{N}C = \bigcup_{i=1}^{k}\overline{N}C_i,$$

respectively.

Usually, $\underline{N}c_i \subseteq \overline{N}c_i$, and we call $BN(c_i) = \overline{N}c_i - \underline{N}c_i$ the boundary region of c_i. We say c_i is δ-neighborhood consistent if $BN(c_i) = \emptyset$. In this case, all the samples in c_i are certainly classified into c_i; otherwise, c_i is not consistent.

It is easy to show that $\overline{N}C = \bigcup_{i=1}^{k}\overline{N}C_i = U$, and $\underline{N}C \subseteq U$. We say the decisions of samples are δ -neighborhood consistent if $\underline{N}C = U$. In this case all the samples are delta-neighborhood consistent. However, a portion of samples are inconsistent in real-world applications; the ratio of consistent samples, computed with $\|\underline{N}C\|/\|U\|$, is defined as the dependency of decision C to features S, denoted by $\gamma_S(C)$, where $\|A\|$ is the cardinality of set A.

The size of neighborhood, controlled by the values of δ, is a parameter to control the granularity when handling classification problems. The coarser the granularity is, the greater the decision boundary region would be. Therefore the classification is more inconsistent in this case. For detailed information, one can refer to literature (Hu et al., 2008).

3. Neighborhood mutual information in metric spaces

Shannon's entropy and mutual information cannot be used to compute relevance between numerical features due to the difficulty in estimating probability density. In this section, we introduce the concept of neighborhood into information theory, and generalize Shannon's entropy for the numerical information.

Definition 1. Given a set of samples $U = \{x_1, x_2, \ldots, x_n\}$ described by numerical or discrete features F, $S \subseteq F$ is a subset of attributes. The neighborhood of sample x_i in S is denoted by $\delta_S(x_i)$. Then the neighborhood uncertainty of the sample is defined as

$$NH_\delta^{x_i}(S) = -\log\frac{\|\delta_S(x_i)\|}{n},$$

and the average uncertainty of the set of samples is computed as

$$NH_\delta(S) = -\frac{1}{n}\sum_{i=1}^{n}\log\frac{\|\delta_S(x_i)\|}{n}.$$

Since $\forall x_i$, $\delta_S(x_i) \subseteq U$, $\|\delta_S(x_i\quad)\|/n \leqslant 1$, so we have $\log n \geqslant NH_\delta(S) \geqslant 0$. $NH_\delta(S) = \log n$ if and only if for $\forall x_i$, $\|\delta_S(x_i)\| = 1$. $NH_\delta(S) = 0$ if and only if for $\forall x_i$, $\|\delta_S(x_i)\| = n$.

Theorem 1. *If $\delta \leqslant \delta'$, $NH_\delta(S) \geqslant NH_{\delta'}(S)$.*

Proof. $\forall x_i \in U$, we have $\delta(x_i) \subseteq \delta'(x_i)$, then $\|\delta(x_i)\| \leqslant \|\delta'(x_i)\|$, we have $NH_\delta(S) \geqslant NH_{\delta'}(S)$. □

Theorem 2. *If $\delta = 0$, then $NH_\delta(S) = H(S)$, where $H(S)$ is Shannon's entropy.*

Proof. If $\delta = 0$, the samples are divided into disjoint $X_1, X_2, \ldots, X_m$, where $\Delta(x_i, x_j) = 0$ if x_i, $x_j \in X_k$. Assumed there are m_i samples in X_i, then $H(S) = -\sum_{i=1}^{m}\frac{m_i}{n}\log\frac{m_i}{n}$. $\delta_S\,(x) = X_k$ if $x \in X_k$ and $\delta = 0$. If $i \neq j$, $X_i \cap X_j = \emptyset$, we have

$$\begin{aligned} NH_\delta(S) &= -\frac{1}{n}\log\frac{\|\delta_S(x_i)\|}{n} \\ &= \sum_{x\in X_1} -\frac{1}{n}\log\frac{\|\delta_S(x)\|}{n} + \cdots + \sum_{x\in X_m} -\frac{1}{n}\log\frac{\|\delta_S(x)\|}{n}. \end{aligned}$$

This leads us to the conclusion that $NH_\delta(S) = H(S)$ if $\delta = 0$. □

Neighborhood entropy is a natural generalization of the Shannon's entropy if features are continuous. As to discrete features, we can define a discrete distance such that $\Delta(x, y) = 0$ if $x = y$; otherwise $\Delta(x, y) = 1$. If $\delta < 1$, the subset $\delta_S(x_i)$ of samples forms the equivalence class $[x_i]$, where $[x_i]$ is the set of samples taking the same feature values with x_i. In this case, the neighborhood entropy equals Shannon entropy.

Definition 2. R, $S \subseteq F$ are two subsets of attributes. The neighborhood of sample x_i in feature subspace $S \cup R$ is denoted by $\delta_{\ R\cup S}(x_i)$, then the joint neighborhood entropy is computed as

$$NH_\delta(R,S) = -\frac{1}{n}\sum_{i=1}^{n}\log\frac{\|\delta_{S\cup R}(x_i)\|}{n}.$$

Especially if R is a set of input variables and C is the classification attributes, we define $\delta_{R\cup C}(x_i) = \delta_R(x_i) \cap c_{x_i}$. Then

$$NH_\delta(R,C) = -\frac{1}{n}\sum_{i=1}^{n}\log\frac{\|\delta_R(x_i) \cap c_{x_i}\|}{n}.$$

Theorem 3. *$NH_\delta(R,S) \geqslant NH_\delta(R)$, $NH_\delta(R,S) \geqslant NH_\delta(S)$.*

Proof. $\forall x_i \in U$, we have $\delta_{S\cup R}(x_i) \subseteq \delta_S(x_i)$ and $\delta_{S\cup R}(x_i) \subseteq \delta_R(x_i)$. Then $\|\delta_{S\cup R}(x_i)\| \leqslant \|\delta_S(x_i)\|$ and $\|\delta_{S\cup R}(x_i)\| \leqslant \|\delta_R(x_i)\|$, therefore $NH_\delta(R,S) \geqslant NH_\delta(R)$, $NH_\delta(R,S) \geqslant NH_\delta(S)$. □

Definition 3. R, $S \subseteq F$ are two subsets of attributes. The conditional neighborhood entropy of R to S is defined as

$$NH_\delta(R|S) = -\frac{1}{n}\sum_{i=1}^{n}\log\frac{\|\delta_{S\cup R}(x_i)\|}{\|\delta_S(x_i)\|}.$$

Theorem 4. *$NH_\delta(R|S) = NH_\delta(R,S) - NH_\delta(S)$*

Proof. $NH_\delta(R,S) - NH_\delta(S)$

$$= -\frac{1}{n}\sum_{i=1}^{n}\log\frac{\|\delta_{S\cup R}(x_i)\|}{n} - \left(-\frac{1}{n}\sum_{i=1}^{n}\log\frac{\|\delta_S(x_i)\|}{n}\right)$$
$$= -\frac{1}{n}\sum_{i=1}^{n}\left(\log\frac{\|\delta_{S\cup R}(x_i)\|}{n} - \log\frac{\|\delta_S(x_i)\|}{n}\right)$$
$$= -\frac{1}{n}\sum_{i=1}^{n}\log\frac{\|\delta_{S\cup R}(x_i)\|}{\|\delta_S(x_i)\|} \quad \square$$

Definition 4. $R, S \subseteq F$ are two subsets of attributes. The neighborhood mutual information of R and S is defined as

$$NMI_\delta(R;S) = -\frac{1}{n}\sum_{i=1}^{n}\log\frac{\|\delta_R(x_i)\| \cdot \|\delta_S(x_i)\|}{n\|\delta_{S\cup R}(x_i)\|}.$$

Theorem 5. *Given two subsets of attributes R and S, $NMI_\delta\ (R;S)$ is the mutual information of these subsets, then the following equations hold:*

(1) *$NMI_\delta(R;S) = NMI_\delta(S;R)$;*
(2) *$NMI_\delta(R;S) = NH_\delta(R) + NH_\delta(S) - NH_\delta(R,S)$;*
(3) *$NMI_\delta(R;S) = NH_\delta(R) - NH_\delta(R|S) \ = NH_\delta(S) - NH_\delta(S|R)$.*

Proof. The conclusions of (1) and (3) are straightforward; here we give the proof of property (2).

(2) $NH_\delta(R) + NH_\delta(S) - NH_\delta(R,S)$

$$= -\frac{1}{n}\sum_{i=1}^{n}\log\frac{\|\delta_R(x_i)\|}{n} - \frac{1}{n}\sum_{i=1}^{n}\log\frac{\|\delta_S(x_i)\|}{n} - \left(-\frac{1}{n}\sum_{i=1}^{n}\log\frac{\|\delta_{R\cup S}(x_i)\|}{n}\right)$$
$$= -\frac{1}{n}\sum_{i=1}^{n}\left(\log\frac{\|\delta_R(x_i)\|}{n} + \log\frac{\|\delta_S(x_i)\|}{n} - \log\frac{\|\delta_{R\cup S}(x_i)\|}{n}\right)$$
$$= -\frac{1}{n}\sum_{i=1}^{n}\left(\log\frac{\|\delta_R(x_i)\|}{n}\frac{\|\delta_S(x_i)\|}{n}\frac{n}{\|\delta_{R\cup S}(x_i)\|}\right)$$
$$= -\frac{1}{n}\sum_{i=1}^{n}\left(\log\frac{\|\delta_R(x_i)\| \cdot \|\delta_S(x_i)\|}{n\|\delta_{R\cup S}(x_i)\|}\right). \quad \square$$

Lemma 1. *Given a set U of samples described by attribute set F, $R \subseteq F$ and C is the decision attribute. $NMI^x_\delta(R;C) = H^x(C)$ if the decision of sample $x \in U$ is δ-neighborhood consistent, where $NMI^x_\delta(R;C) = -\log\frac{\|\delta_R(x)\|\cdot\|c_x\|}{n\|\delta_{R\cup C}(x)\|}$, $H^x(C) = -\log\frac{\|c_x\|}{n}$.*

Proof. $\delta_{R\cup C}(x) = \delta_R(x) \cap c_x$, and we have that $\delta_R(x) \subseteq c_x$ if x is consistent. In this case $\delta_{\ R\cup C}(x) = \delta_R(x)$. Then

$$-\log\frac{\|\delta_R(x)\| \cdot \|c_x\|}{n\|\delta_{R\cup C}(x)\|} = -\log\frac{\|\delta_R(x)\| \cdot \|c_x\|}{n\|\delta_R(x)\|} = -\log\frac{\|c_x\|}{n}. \quad \square$$

Theorem 6. *Given a set of samples U described by the attribute set F, $R \subseteq S$ and C is the decision attribute. $NMI_\delta\ (R;C) = H(C)$ if the decisions of samples in feature subspace R are δ-neighborhood consistent.*

Proof. As the decisions of samples in feature subspace are consistent, the decision of each sample is consistent. For $\forall x_i \in U$, $NMI^{x_i}_\delta(R;C) = H^{x_i}(C)$. So $\sum_{i=1}^{n} NMI^{x_i}_\delta(R;C) = \sum_{i=1}^{n} H^{x_i}(C)$.

$$\sum_{i=1}^{n} NMI^{x_i}_\delta(R;C) = NMI_\delta(R;C); \qquad \sum_{i=1}^{n} H^{x_i}(C) = H(C).$$

We get the conclusion $NMI_\delta(R;C) = H(C)$. □

Theorem 6 shows that the mutual information between features R and decision C equals to the uncertainty quantity of decision if the classification is consistent with respect to the knowledge of R. There is not any uncertainty in classification if attributes R is known. Moreover, we also know by Lemma 1 that the mutual information between R and C with respect to sample x is the uncertainty of x in classification if its decision is consistent. With Lemma 1 and Theorem 6, we not only distinguish whether all samples in classification learning are consistent, but also know which samples are consistent although the decision is not consistent as a whole. In practice it is often that just part of consistent samples is consistent. It is useful to find these consistent patterns for understanding the task at hand.

4. Strategies for selecting features

4.1. Axiomatization of feature selection

Neighborhood mutual information measures the relevance between numerical or nominal variables. It is also shown that the neighborhood entropy will degenerate to the Shannon's entropy if the features are nominal, thus neighborhood mutual information will reduce to the classical mutual information. Mutual information is widely used in selecting nominal features. We extend these algorithms to select numerical and nominal features by computing relevance with neighborhood mutual information.

As discussed in the theory of rough sets, a "sound" subset of features should be sufficient and necessary. Sufficiency requires the selected features have the same capability in describing the decision as the whole set of features; necessary shows no superfluous features in the selected subset (Yao & Zhao, 2008).

Axiom 1.1 (*Sufficiency condition*). Given a dataset described by feature set F and decision variable C, a feature subset $S \subseteq F$ is said to be a sufficient feature subset if $r_F(C) = r_S(C)$, where $r_S(C)$ is the dependency of classification C on features S.

Axiom 1.2 (*Indispensability condition*). Given a dataset described by feature set F and decision C, S is a feature subset. $f \in S$ is said to be indispensable if $r_F(C) > r_{S-f}(C)$.

These axioms offer an axiomatic approach to feature subset selection. In rough sets, the subset of features satisfying the sufficiency and indispensability conditions is called a relative reduct in rough set theory. Given a training dataset, there are a lot of relative reducts in some applications (Hu, Yu, Xie, & Li, 2007). The one with the minimal features is favored according to the principle of Occam's razor. It is obvious that a minimal subset of features does not necessarily generate a minimal description of the classification data. The size of the description is related with the number of feature values. The above axioms fail to reflect this fact.

Based on information theory, Wang, Bell and Murtagh introduced the second axiomatic approach to feature subset selection (Bell & Wang, 2000; Wang et al., 1999).

Axiom 2.1 (*Preservation of learning information*). For a given dataset described by features F and decision variable C, the expected feature subset, S, is a sufficient feature subset if $MI(F; C) = MI(S; C)$.

Axiom 2.2 (*Minimum encoding length*). Given a dataset by features F and decision $C, \mathbb{S}$ is a set of sufficient feature subsets. The one, $S \in \mathbb{S}$, which minimizes the joint entropy $H(S,C)$ should be favored with respect to its predictive capability.

Axioms 2.1 and 2.2 give an axiomatic description of a good subset of features based on information theory and the principle of Occam's razor. In fact, as to a consistent classification problem, we can easily get the following property (Hu, Yu, & Xie, 2006): If $r_F(C) = r_S(C)$, we have $MI(F; C) = MI(S; C)$.

We consider both dependency and mutual information are measures of relevance between features, then the above two axiomatic approaches require that the relevance between the reduced subsets of features does not decrease.

The difference comes from the second term of two approaches. In the framework of rough sets, the reduct with minimal features is preferred, while the features minimizing the joint entropy are preferred according to information theory. Entropy was viewed as a measure of granularity of partitioning objects based on the values of features (Qian, Liang, & Dang, 2009; Yu, Hu, & Wu, 2007). Minimizing the joint entropy leads to select a subset of features which maximizes the granularity of partition derived jointly with the features and the decision variable.

It is difficult to apply this description to the problem of numerical feature selection. Now we present the third axiomatic system that is suitable for both nominal and numerical feature subset selection.

Axiom 3.1 (*Preservation of learning information under granularity δ*). Given a dataset described by features F and decision variable C, the expected feature subset, S, is sufficient if $NMI_\delta(S; C) = NMI_\delta(F; C)$ with respect to granularity δ.

Axiom 3.2 (*Minimum encoding length under granularity δ*). Given a dataset by features F and decision C, $\mathbb{S}$ is a set of sufficient feature subsets. The one, $S \in \mathbb{S}$, which minimizes the joint entropy $NH_\delta(S,C)$ should be favored with respect to its predictive capability under granularity δ.

It is notable that Axiom 3 gives a multi-granular way to describe the classification power of a set of numerical features because δ can be considered as a variable. Multi-granular analysis can be conducted in discussing a classification problem. We have the following properties of monotonicity.

Theorem 7. (Type-1 Monotonicity). *Given a consistent classification problem described by features F and decision C, $S \subseteq F$ is a sufficient feature subset with respect to granularity δ. If $S \subseteq R \subseteq F$, R is also a sufficient feature subset.*

Proof. As we know $NMI_\delta(F; C) = H(C)$ if the classification problem is consistent, and $NMI_\delta(S; C) = NMI_\delta(F; C)$, then $NMI_\delta(S; C) = H(C)$.

$S \subseteq R$, so $NMI_\delta(R; C) \geqslant NMI_\delta(S; C)$.

$H(C) \geqslant NMI_\delta(R; C)$. Finally $NMI_\delta(R; C) = H(C)$, R is a sufficient feature subset. □

Theorem 8. (Type-2 Monotonicity). *Given a consistent classification problem described by features F and decision C, $S \subseteq F$. $0 \leqslant \delta_1 \leqslant \delta_2$, S is a sufficient feature subset under granularity δ_1 if S is a sufficient feature subset under granularity δ_2.*

Proof. S is a sufficient feature subset, so we have $NMI_{\delta_2}(S; C) = NMI_{\delta_2}(F; C) = H(C)$. This reflects the classification problem in feature space S is consistent under granularity δ_2. As $0 \leqslant \delta_1 \leqslant \delta_2$, the classifications in S and F under granularity δ_1 are consistent if the classification in feature space S under granularity δ_2 is consistent. So $NMI_{\delta_1}(S; C) = H(C)$ and $NMI_{\delta_1}(F; C) = H(C)$. We have $NMI_{\delta_1}(S; C) = NMI_{\delta_1}(F; C)$. □

4.2. Feature selection algorithms

The axiomatic approaches set a goal for feature subset selection. That is, the expected subset S of features should be sufficient and with the minimal joint entropy $NH_\delta(S,C)$. A straightforward way is to exhaustively check the subsets of features to find an expected subset. However, this is not feasible even given a moderate size of candidate features due to the exponential complexity.

Some efficient algorithms were developed to overcome this problem. Battiti in Battiti (1994) and Peng et al. (2005) discussed two criteria, named Max-Relevance (MR), Minimal-Redundancy and Max-Relevance (mRMR), respectively. We here will introduce two new criteria named Maximal-Dependency (MD) and Minimal-redundancy and Maximal-Dependency (mRMD). Furthermore, we will offer a new interpretation in terms of neighborhood mutual information.

Intuitively, features of greater relevance with decision should provide more information for classification. Therefore, the best feature should be the one of the greatest mutual information. This strategy is called maximal relevance criterion (Max-Relevance, MR). Formally, Max-Relevance criterion can be written as the following formulation:

$$\max D(S,C), \quad D = \frac{1}{\|S\|} \sum_{f_i \in S} NMI_\delta(f_i; C).$$

In essence the MR criterion is a feature selection algorithm based on ranking. We rank the features in the descending order according to the mutual information between single features and decision, and then select the first k features, where k has been specified in advance.

It is well known that ranking based algorithm cannot remove redundancy between features because this algorithm neglects the relevance between input variables. Sometimes, the redundancy between features is so great that deleting some of them would not reduce the classification information of the original data. In this case, we should select a subset of features with the minimal redundancy condition. That is

$$\min(R),\quad R=\frac{1}{\|S\|^2}\sum_{f_i,f_j\in S} NMI_\delta(f_i;f_j).$$

Then we get a new criterion, called minimal-Redundancy-Maximal-Relevance (mRMR), by combining the above two constraints

$$\max \Phi(D,R),\quad \Phi = D\,-\,\beta\,R,$$

where the parameter β is used to regulate the relative importance of the mutual information between the features and the decision.

mRMR computes the significance of each feature one by one, and ranks the features according their significances in the descending order. Then some classification algorithm is introduced to check the best k features with respect to the classification performance, where $k = 1,\ldots,N$, N is the number of all candidate features.

Another alternative of selection criterion is to maximize the joint relevance between features and decision with a greedy algorithm; as a by-product, the redundancy among features might be reduced. This criterion is called Maximal-Dependency (MD). In each round, we select a feature which produces the maximal increase of joint mutual information, formally written as

$$\max_{f\in F-S}\Psi(f,S,C),\quad \Psi(f,S,C)=NMI_\delta(S\cup\{f\};C)-NMI_\delta(S;C).$$

It is known that

$$NMI_\delta(S;C)=-\frac{1}{n}\sum_{i=1}^{n}\log\frac{\|\delta_S(x_i)\|\|\delta_C(x_i)\|}{n\|\delta_S(x_i)\cap\delta_C(x_i)\|},\ \text{then}$$

$$\begin{aligned}NMI_\delta(S\cup\{f\};C)-NMI_\delta(S;C)&=-\frac{1}{n}\sum_{i=1}^{n}\log\frac{\|\delta_S(x_i)\|\|\delta_C(x_i)\|}{n\|\delta_S(x_i)\cap\delta_C(x_i)\|}+\frac{1}{n}\\&\quad\times\sum_{i=1}^{n}\log\frac{\|\delta_{S\cup\{f\}}(x_i)\|\|\delta_C(x_i)\|}{n\|\delta_{S\cup\{f\}}(x_i)\cap\delta_C(x_i)\|}\\&=-\frac{1}{n}\sum_{i=1}^{n}\\&\quad\times\log\frac{\|\delta_{S\cup\{f\}}(x_i)\|\|\delta_S(x_i)\cap\delta_C(x_i)\|}{\|\delta_{S\cup\{f\}}(x_i)\cap\delta_C(x_i)\|\|\delta_S(x_i)\|}\end{aligned}$$

If we compute the neighborhood with infinite norm based distance, $\delta_{S\cup\{f\}}(x_i) = \delta_S(x_i)\cap\delta_f(x_i)$. In this case,

$$\begin{aligned}NMI_\delta(S\cup\{f\};C)-NMI_\delta(S;C)&=-\frac{1}{n}\sum_{i=1}^{n}\\&\quad\times\log\frac{\|\delta_S(x_i)\cap\delta_f(x_i)\|\|\delta_S(x_i)\cap\delta_C(x_i)\|}{\|\delta_S(x_i)\cap\delta_f(x_i)\cap\delta_C(x_i)\|\|\delta_S(x_i)\|},\end{aligned}$$

We set

$$p^{x_i}(C|S)=\frac{\|\delta_S(x_i)\cap\delta_C(x_i)\|}{\|\delta_S(x_i)\|}\quad\text{and}$$

$$p^{x_i}(C|S\cup\{f\})=\frac{\|\delta_S(x_i)\cap\delta_f(x_i)\cap\delta_C(x_i)\|}{\|\delta_S(x_i)\cap\delta_f(x_i)\|},$$

then

$$NMI_\delta(S\cup\{f\};C)-NMI_\delta(S;C)=-\frac{1}{n}\sum_{i=1}^{n}\log\frac{p^{x_i}(C|S)}{p^{x_i}(C|S\cup\{f\})}.$$

This conclusion shows that maximizing the function $\Psi(f,S,C)$ translates into adding feature f which leads to the maximal increase of classification probability. This feature is obviously expected for classification. Here we implicitly estimate the probability and class probability with the samples in neighborhoods in evaluating features with neighborhood mutual information. Imprecise estimation would not have great influence on the finally result as we just obtain the best feature in each round.

Fig. 1 shows the neighborhoods of sample x_i in different feature subspaces. $\delta_C(x_i)$, $\delta_S(x_i)$ and $\delta_f(x_i)$ are the neighborhoods of sample x_i in terms of decision variable, the currently subset S of selected features and a new candidate feature f. In Fig. 1 (1), $\delta_S(x_i)\not\subset\delta_C(x_i)$. After adding f into S, $\delta_{S\cup\{f\}}(x_i) = \delta_S(x_i)\cap\delta_f(x_i)$ is not contained by $\delta_C(x_i)$ yet. This shows sample x_i is not consistent in subspace S and $S\cup\{f\}$. However, we see in Fig. 1 (2) that although x_i is not consistent in subspace S, it is consistent in $S\cup\{f\}$. In this case,

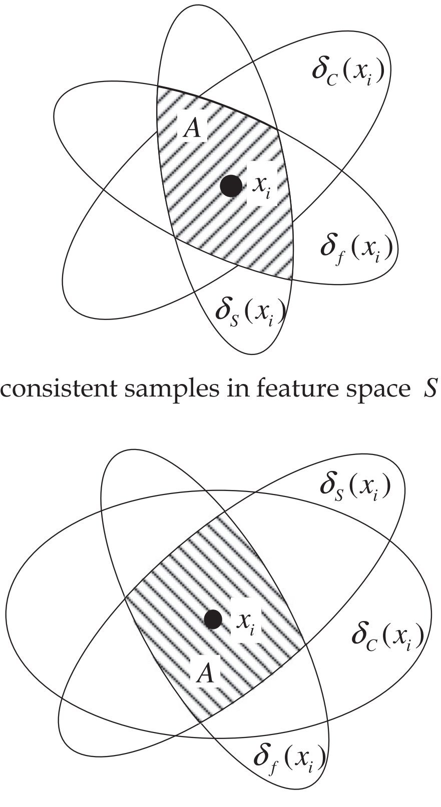
$$NMI_\delta^{x_i}(S\cup\{f\};C)=H^{x_i}(C)>NMI_\delta^{x_i}(S;C).$$

MD is a locally optimal algorithm which selects the best feature in current rounds. However, the selected feature might be not globally optimal. Moreover, this algorithm overlooks the redundancy between features; we thus give a minimal-Redundancy-Maximal-Dependency algorithm (mRMD):

$$\max\Theta(S,C),\quad \Theta=NMI_\delta(S;C)-\frac{\beta}{\|S\|^2}\sum_{f_i,f_j\in S}NMI_\delta(f_i;f_j).$$

It is worth noting that the ideal features should be globally maximal relevant, namely maximizing $NMI_\delta(S;C)$. However, searching the globally optimal features is NP-hard. All the above four criteria contribute to approximate solutions.

Now we discuss the complexity of these algorithms. Given N candidate features, we need compute the relevance between N features and decision variable. So the time complexity is $O(N)$. As to the mRMR criterion, assumed k features have been selected, then we should compute the relevance between the $N-k$ remaining



(1) Inconsistent samples in feature space $S\bigcup\{f\}$

(2) Consistent samples in feature space $S\bigcup\{f\}$

Fig. 1. Neighborhood inconsistent and consistent samples.

features and decision. Besides, we also require computing the relevance between the $N-k$ remaining features and the k selected features. The computational complexity in this round is $N-k+(N-k)\times k$. The total complexity is $\sum_{k=1}^{N} N-k+(N-k)\times k$. Therefore, the time complexity of mRMR is $O(N^3)$. As to MD, assumed k features, denoted by S_k, have been selected in the kth current round, then we should compute the joint mutual information of S_k and the remaining $N-k$ features. The complexity here is $O(N-k)$. The total complexity is $O\left(\sum_{k=1}^{N} N-k\right)=O\left(N^2\right)$ in the worst case. The complexity of mRMD is the same as mRMR. In summary, computational complexity of MR is linear, MD is quadratic, whereas mRMR and mRMD are cubic.

5. Experimental analysis

In this section, we will first show the properties of neighborhood mutual information. Then we compare the neighborhood mutual information based MD feature selection algorithm with neighborhood rough sets based algorithm (NRS) (Hu et al., 2008), correlation based feature selection (CFS) (Hall, 2000), consistency based algorithm (Dash & Liu, 2003) and FCBF (Yu & Liu, 2004). Finally, the effectiveness of mRMR, MD and mRMD is discussed.

5.1. Properties of neighborhood mutual information

First we use data Iris to reveal the effectiveness of neighborhood mutual information. The data set contains 3 classes (Setosa, Versicolour, Virginica) of 50 instances each, where each class refers to a type of iris plant. One class is linearly separable from the two others; the latter are not linearly separable from the others. Each sample is described by four numerical features: sepal length (SL), sepal width (SW), petal length (PL), and petal width (PW). The scatter plots of samples in feature-pair subspaces are shown in Fig. 2.

It is easy to find that two classes of samples in subspace SL-SW are nearly overlapped; however, there are just a few inconsistent samples in PL-PW. Intuitively, classification accuracy in PL-PW would be better than SL-SW. Therefore it is expect that the mutual information between PL-PW and classification should be greater than that between SL-SW and classification.

We normalize each feature into the unit interval [0, 1], and set $\delta = 0.1$. The neighborhood mutual information of each feature-pair is given in Table 1. Moreover, we also give the 10-fold cross validation classification accuracy based on CART learning algorithm in Table 2.

From Table 1, we learn that the subset of features PW and PL gets the greatest NMI. Correspondingly, these two features also produce the highest classification accuracy. We compute the correlation coefficient between matrices of NMI and classification accu-

Table 1
NMI of each feature-pair.

	SL	SW	PL	PW
SL	0.7207	1.0463	1.3227	1.4048
SW	1.0463	0.3743	1.3314	1.4149
PL	1.3227	1.3314	1.3543	1.4549
PW	1.4048	1.4149	1.4549	1.4125

Table 2
10-CV accuracy with CART.

	SL	SW	PL	PW
SL	0.7200	0.7000	0.9467	0.9534
SW	0.6867	0.4800	0.9467	0.9400
PL	0.9533	0.9467	0.9467	0.9667
PW	0.9533	0.9400	0.9667	0.9534

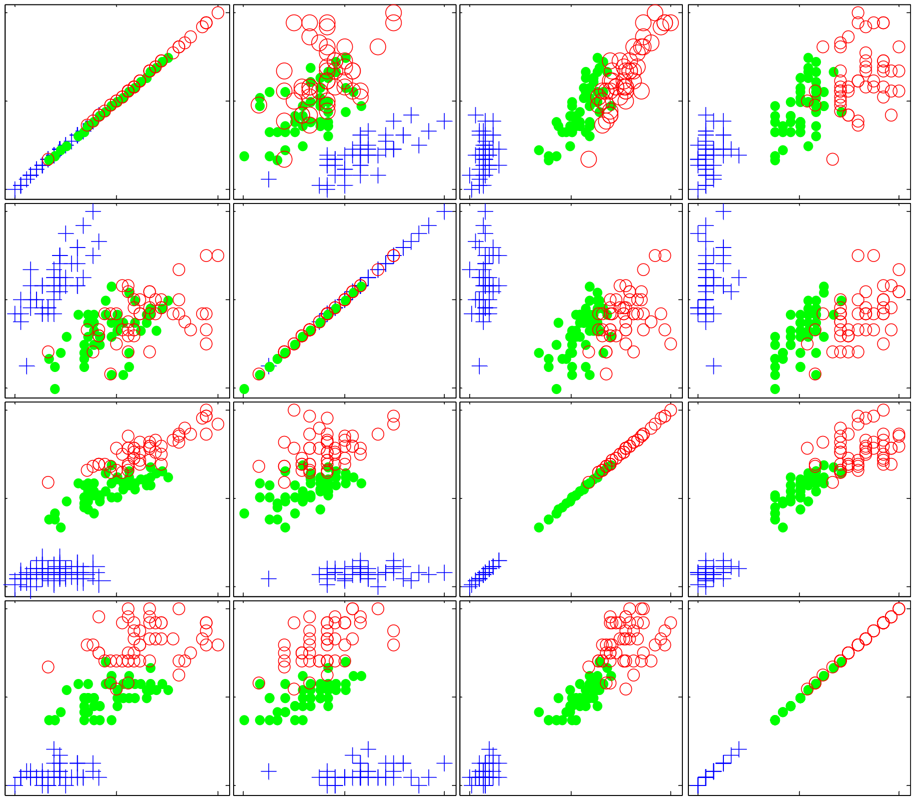


Fig. 2. Scatter plots of samples in feature-pair subspaces.

racies. The value of coefficient is equal to 0.95. It shows that NMI is effective for estimating classification performance.

In computing NMI, we had to specify a value of parameter δ. Now we discuss its influence on NMI. According to Theorem 1, we know if $\delta_1 \leqslant \delta_2$, $NH_{\delta_1}(S) \geqslant NH_{\delta_2}(S)$. However, we do not get a similar conclusion that $NMI_{\delta_1}(S;C) \geqslant NMI_{\delta_2}(S;C)$ if $\delta_1 \leqslant \delta_2$. We compute NMI of single features when $\delta = 0.1, 0.13, 0.16, \ldots, 0.4$ and the features were normalized into [0,1]. This experiment is conducted on data *heart* and *wine*. There are 7 nominal features and 6 continuous variables in *heart*. As to the nominal features, their values are recoded with a set of integer numbers, whereas the continuous features take values in [0,1]. In this case, NMI will not change when δ varies in interval [0,1). Moreover, NMI is equivalent to mutual information in the classical information theory. The samples in data wine are described with 13 continuous features. Fig. 3 presents the NMI computed with different δ of each feature. The curves from the top down are computed with δ from 0.1 to 0.4, respectively.

From Fig. 3, we observe that NMI of features varies with parameter δ. In a fine granularity, namely, δ is small, the classification information provided by a numeric feature is more than that in a coarse granularity because the classification boundary is small if the problem is analyzed at a fine granularity. As a result, given a certain feature, NMI becomes smaller when the values of δ increase. Moreover, it is worth noting that the order of feature significances can not be retained when granularity changes. Then the sets of selected features would be different. For example, NMI of feature 12 is greater than that of feature 13 if $\delta = 0.4$. However, feature 13 outperforms feature 12 if $\delta = 0.19$.

Fig. 4 presents the classification accuracy of the selected features with respect to the size of neighborhood (heart and wine data sets). We can see that the accuracies do not change much in terms of the results produced by CART, LSVM, RBFSVM and KNN when δ assumes value from 0.02 to 0.4 with step 0.02. According to observations made in Hu et al. (2008), δ should take value in [0.1,0.2]. In the following, if not specified, $\delta = 0.15$.

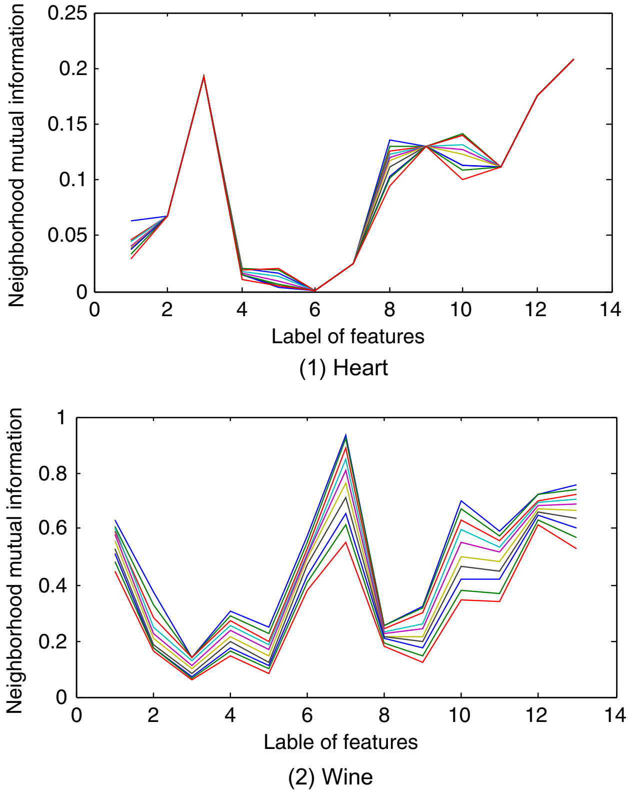


Fig. 3. Neighborhood mutual information between each feature and decision.

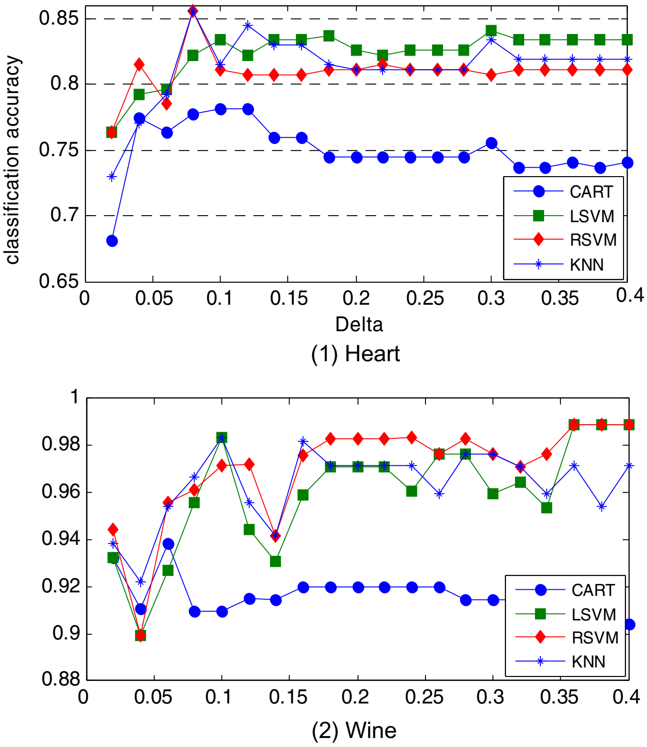


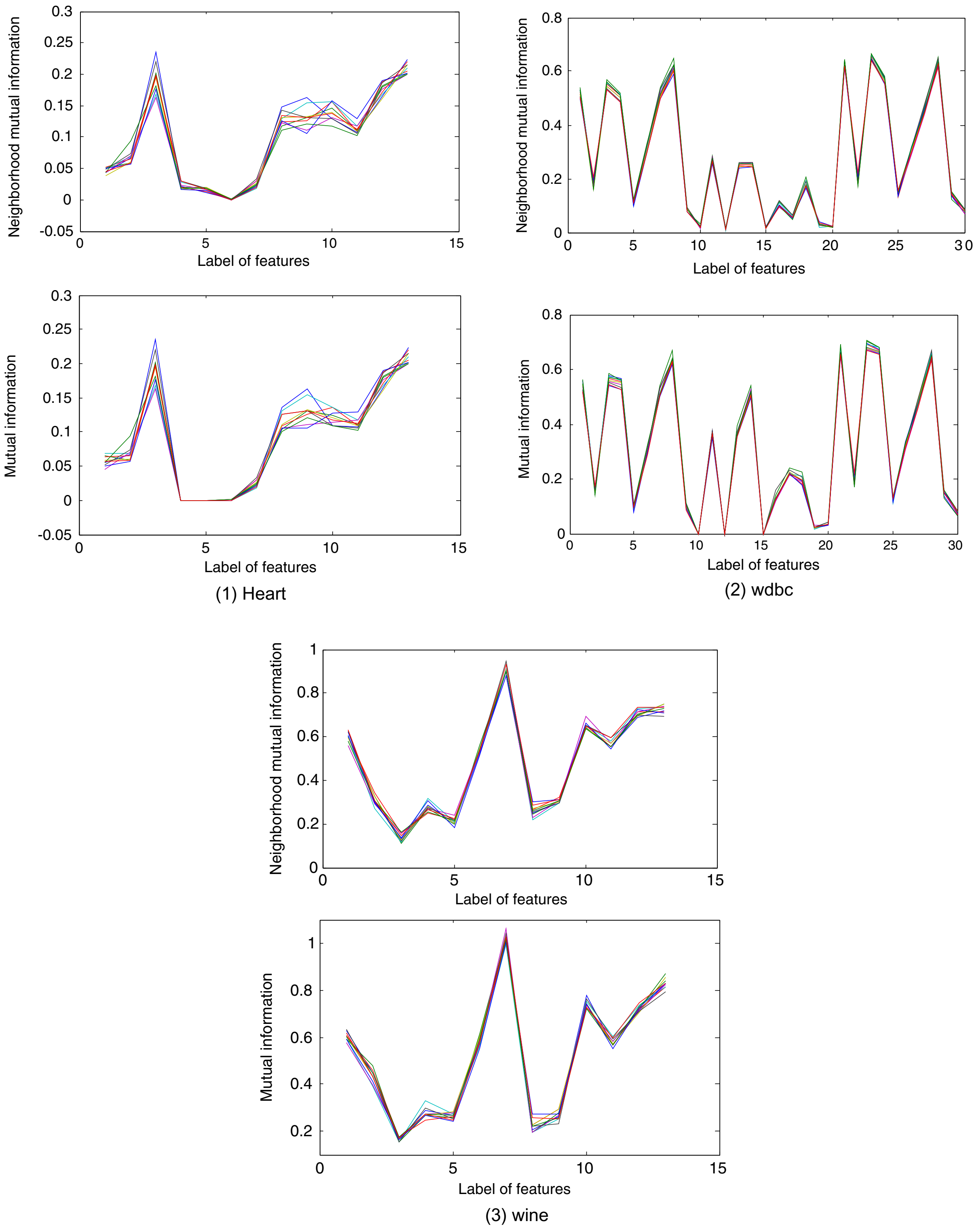
Fig. 4. Classification accuracy of features selected with different δ.

Moreover, we can also illuminate the effectiveness by comparing neighborhood mutual information and mutual information. In order to compute the mutual information of continuous features, we introduce a discretizing algorithm to transform these features into discrete ones (Fayyad and Irani, 1993). We introduce a 10-CV like technique to compute the NMI and MI. That is, the samples are divided into 10 subsets; nine of them are combined to compute the mutual information between single features and decision. After ten rounds, we can get 10 estimates of features. By this way, we can study the stability of the estimation (Kalousis, Prados, and Hilario, 2007).

Three data sets (Heart, WDBC, Wine) are tested. Data *Wisconsin Diagnostic Breast Cancer* (WDBC) is a widely used one in machine learning research. WDBC is a binary classification problem, where 569 samples are characterized with 30 numerical features.

NMI and MI of each feature are given in Fig. 5. Surprisingly, we see that NMI and MI computed with continuous features and their discrete ones are very similar. There are just some little different points between NMI and MI. However, we can also find some information is changed in discretization. As to data wine shown in Fig. 5 (1), we see the mutual information quantities of features 4, 5 and 6 are different before discretization. However, they are the same if the numerical features are discretized. It shows that the difference of features is lost in discretization. As to WDBC, MI quantities of features 13, 14 and 17 change much. Before discretization, feature 18 is better than feature 17. However, feature 17 outperforms feature 18 after discretization.

Moreover, NMI and MI are stable. Although we alter samples in computing NMI and MI, the information quantities do not vary much in each computation. This property of NMI and MI is very important as one usually expects that the same features should be obtained even though sampling is different. For a given classification task, if multiple learning data sets are gathered, we naturally expect we will get the same features from these sets of samples for a feature selection algorithm.



(1) Heart

(2) wdbc

(3) wine

Fig. 5. NMI and MI between each feature and decision.

The above experiments show that NMI is an effective substitute of mutual information for computing relevance between numerical features without discretization.

5.2. Comparison of NMI with related algorithms

In order to compare NMI based feature selection algorithms with some classical techniques, 15 databases are downloaded from UCI Repository of machine learning databases (Blake and Merz, 1998). The description of data is presented in Table 3. The sizes of databases vary from 155 to 20000, and the numbers of candidate features vary from 13 to 649. Moreover, we gathered a data set, named vibration, which describes a problem of vibration diagnosis for gas engine. We acquired the wave data samples with different faults, and then introduced wavelet techniques to extract 72 features from these waves. The number of candidate features is given

Table 3
Experimental data description and the numbers of selected with different algorithms.

ID	Data	Samples	Classes	Features	NMI	CFS	Consistency	FCBF	NRS
1	german	1000	2	3/17	12	5	11	5	11
2	heart	270	2	5/ 8	8	7	10	6	9
3	hepatitis	155	2	6/13	6	7	4	7	7
4	horse	368	2	7/15	7	8	4	8	8
5	iono	351	2	32/0	8	14	7	4	9
6	letter	20000	26	0/16	9	11	11	11	16
7	m-feature	2000	10	649/0	7	8	11	4	9
8	mushroom	8124	2	0/22	3	5	5	5	5
9	sick	2800	2	5/24	14	4	8	6	23
10	segmentation	2310	7	16/2	8	7	10	5	14
11	sonar	208	2	60/0	6	19	14	10	7
12	spam	4601	2	57/0	24	15	25	14	16
13	vibration	414	5	72/0	10	18	7	10	11
14	wdbc	569	2	30/0	6	11	8	7	12
15	wine	178	3	13/0	5	11	5	10	6
16	wpbc	198	2	33/0	6	2	2	2	7
Average		–	–	69.3	8.7	9.5	8.9	7.1	10.6

as continuous/discrete ones. Among 16 data sets, two are completely discrete, eight are completely continuous and the rest 6 are heterogeneous. All the continuous features are transformed to interval [0,1] in preprocessing, while the discrete features are coded with a sequence of integers. As CFS, consistency and FCBF cannot deal with numerical features directly. We employ a discretization algorithm to transform the numerical features into discrete one (Fayyad and Irani, 1993).

We compare NMI based MD feature selection algorithm with neighborhood rough sets based algorithm (NRS) (Hu et al., 2008), correlation based feature selection (CFS) (Hall, 2000), consistency based algorithm (Dash and Liu, 2003) and FCBF (Yu and Liu,

Table 4
Classification accuracies (%) of features selected with different algorithms based on CART.

Data	Raw	NMI	NRS	CFS	Consistency	FCBF
german	69.9 ± 3.5	71.4 ± 3.6	70.6 ± 5.2	69.7 ± 5.0	68.6 ± 4.6	69.8 ± 4.9
heart	74.1 ± 6.3	78.1 ± 8.1	75.9 ± 7.7	77.0 ± 6.9	76.3 ± 6.3	77.4 ± 7.1
hepatitis	91.0 ± 5.5	86.8 ± 6.5	90.3 ± 4.6	93.0 ± 7.1	89.8 ± 8.6	93.0 ± 7.1
horse	95.9 ± 2.3	89.4 ± 4.8	88.9 ± 5.6	95.9 ± 1.9	95.4 ± 4.3	95.9 ± 1.9
iono	87.6 ± 6.9	93.2 ± 3.7	88.4 ± 6.6	88.7 ± 7.1	88.4 ± 6.6	87.8 ± 7.0
letter	82.3 ± 1.2	86.2 ± 0.9	86.9 ± 1.0	86.6 ± 1.0	86.7 ± 1.0	86.7 ± 1.1
m-feature	93.3 ± 1.7	92.4 ± 2.0	91.4 ± 1.9	45.5 ± 2.6	43.7 ± 3.2	41.9 ± 3.7
mushroom	96.4 ± 9.9	96.0 ± 9.8	96.4 ± 9.9	95.6 ± 9.7	96.7 ± 9.9	95.6 ± 9.7
sick	98.5 ± 1.2	98.5 ± 1.2	98.5 ± 1.1	95.1 ± 1.4	98.1 ± 1.0	95.2 ± 1.3
segmentation	95.6 ± 2.8	94.8 ± 3.7	95.0 ± 3.6	96.1 ± 2.1	95.9 ± 2.7	95.1 ± 2.2
sonar	72.1 ± 13.9	77.4 ± 4.1	69.7 ± 13.2	70.7 ± 14.1	75.5 ± 10.5	70.6 ± 12.1
spam	90.6 ± 3.3	89.3 ± 4.2	85.0 ± 6.8	90.5 ± 3.4	88.9 ± 3.4	90.9 ± 3.2
Vibration	86.5 ± 6.7	87.1 ± 4.7	79.0 ± 6.6	88.4 ± 7.1	91.6 ± 5.4	91.9 ± 4.4
wdbc	90.5 ± 4.6	91.8 ± 3.4	94.0 ± 4.2	92.8 ± 4.8	93.2 ± 4.1	94.0 ± 4.6
wine	89.9 ± 6.4	91.0 ± 6.0	91.5 ± 6.1	89.9 ± 6.3	94.4 ± 3.7	90.4 ± 6.5
wpbc	70.6 ± 7.5	66.6 ± 10.6	70.7 ± 8.4	72.7 ± 10.6	72.7 ± 10.6	72.7 ± 10.6
Average	86.6	86.9	85.8	84.3	84.7	84.3

Table 5
Classification accuracies (%) of features selected with different algorithms based on KNN.

Data	Raw	NMI	NRS	CFS	Consistency	FCBF
german	69.4 ± 2.2	73.6 ± 6.0	70.5 ± 3.4	70.2 ± 4.4	71.6 ± 4.6	70.2 ± 4.4
heart	81.9 ± 6.1	83.3 ± 6.1	83.0 ± 7.0	83.0 ± 4.7	84.1 ± 7.6	83.3 ± 5.9
hepatitis	87.2 ± 5.9	84.5 ± 4.4	90.2 ± 8.6	87.5 ± 7.5	89.8 ± 7.3	87.5 ± 7.5
horse	89.9 ± 4.2	88.6 ± 4.3	89.9 ± 3.4	91.6 ± 4.9	88.3 ± 4.6	91.6 ± 4.9
iono	84.1 ± 5.8	82.7 ± 6.2	88.0 ± 2.2	86.1 ± 7.6	88.0 ± 2.2	88.7 ± 4.9
letter	95.5 ± 0.5	94.3 ± 1.3	93.4 ± 0.9	95.0 ± 0.5	95.0 ± 0.5	95.0 ± 0.4
m-feature	97.7 ± 1.1	96.3 ± 1.4	95.2 ± 1.6	38.6 ± 4.3	35.3 ± 3.9	31.0 ± 2.1
mushroom	94.6 ± 11.2	93.2 ± 14	95.6 ± 10.3	95.7 ± 9.7	95.7 ± 10.1	95.7 ± 9.7
sick	95.8 ± 0.9	95.9 ± 1.0	95.9 ± 1.0	95.4 ± 0.7	96.9 ± 0.9	95.4 ± 0.7
segmentation	94.5 ± 3.5	95.3 ± 3.3	94.5 ± 3.5	94.8 ± 3.7	94.2 ± 3.8	95.1 ± 3.1
sonar	81.3 ± 6.1	81.3 ± 6.5	77.4 ± 9.8	81.7 ± 5.4	86.6 ± 6.8	81.2 ± 7.6
spam	88.2 ± 3.5	87.1 ± 2.7	80.2 ± 4.9	91.3 ± 3.9	87.8 ± 4.0	91.0 ± 3.6
Vibration	94.5 ± 3.0	90.6 ± 5.2	90.0 ± 5.0	92.2 ± 3.1	90.3 ± 3.1	93.2 ± 2.4
wdbc	96.8 ± 2.3	96.1 ± 2.3	95.1 ± 2.3	96.8 ± 2.3	95.1 ± 3.0	95.8 ± 2.0
wine	95.4 ± 4.6	98.3 ± 2.7	97.6 ± 4.2	97.2 ± 3.0	96.0 ± 4.6	96.6 ± 2.9
wpbc	75.7 ± 9.1	73.7 ± 6.4	75.8 ± 5.6	73.1 ± 2.4	73.0 ± 2.4	73.0 ± 12.4
Average	88.9	88.4	88.3	85.6	85.5	85.3

2004). NRS evaluates the features with a function called dependency, which is the ratio of consistent samples over the whole learning samples; CFS first discretizes continuous features and then uses symmetric uncertainty to estimate the relevance between discrete features. The significance of a set S of features is computed as

$$SIG_S = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}},$$

where $\overline{r_{cf}}$ is the average of relevance between decision and features, $\overline{r_{ff}}$ is the average of relevance between features, k is number of features in the subset.

Consistency based algorithm was introduced by Dash and Liu (2003), where consistency is the ratio of samples correctly recognized according to the majority voting technique. Among the samples with the same feature values, some of them come from the majority class, while others belong to the minority classes. According to the majority voting technique, only the samples with the minority classes will not be correctly classified. Dash and Liu computed the ratio of samples correctly classified as consistency.

FCBF also employed symmetrical uncertainty to evaluate features. However, this algorithm introduced a new search technique, called Fast Correlation-Based Filter (FCBF). The algorithm selects predominant features and deletes those highly correlating with predominant features. If there are many redundant features, the algorithm would be very fast.

The selected features are validated with 10-fold-cross-validation based on two popular classification algorithms: CART and KNN ($K = 5$).

The numbers of selected features obtained when running different algorithms are presented in Table 3. From the experimental results, we can observe that most of features are removed. FCBF averagely selects the least features; moreover it also gets the smallest subsets of features for 7 data sets among 5 algorithms, while NMI produces 4 smallest features. As a whole, NMI averagely gets 8.7 features for 16 databases. CFS, consistency and NRS select more features than NMI. NRS gets 10.6 features. Roughly speaking, two more features are selected by NRS.

Now we analyze the performance of these selected features. Average accuracies and standard deviations are shown in Tables 4 and 5, respectively.

First, we can conclude that although most of the candidate features are removed from the raw data, the classification accuracies do not decrease too much. *t*-test shows at the 5% significance level, the average accuracies derived from the raw datasets are the same as the ones produced with NMI reduced datasets with respect to CART and KNN. It shows that NMI is an effective measure for feature selection. With respect to CART learning algorithm, the average accuracy is 86.9% for NMI, while 85.8% for NRS. The average classification accuracy reduced 1.1%. However, NMI is a little worse than CFS, consistency and FCBF in terms of CART and KNN. This is caused by different search strategies. Tables 5 and 6 will show the effective of NMI if it is combined with mRMR.

There is a question in feature selection. Specifically, are all the selected features useful for classification learning? As the filter based algorithms evaluate the quality of features with classifier independent measures, the applicability of these features should be validated with the final classification algorithms. One technique to check applicability of selected features is to add features for learning one by one in the order that the features are selected. Then we get a set of nested subsets of features. We compute the classification performances of these subsets. The variation of classification accuracies of the features selected with NMI are given in Fig. 6.

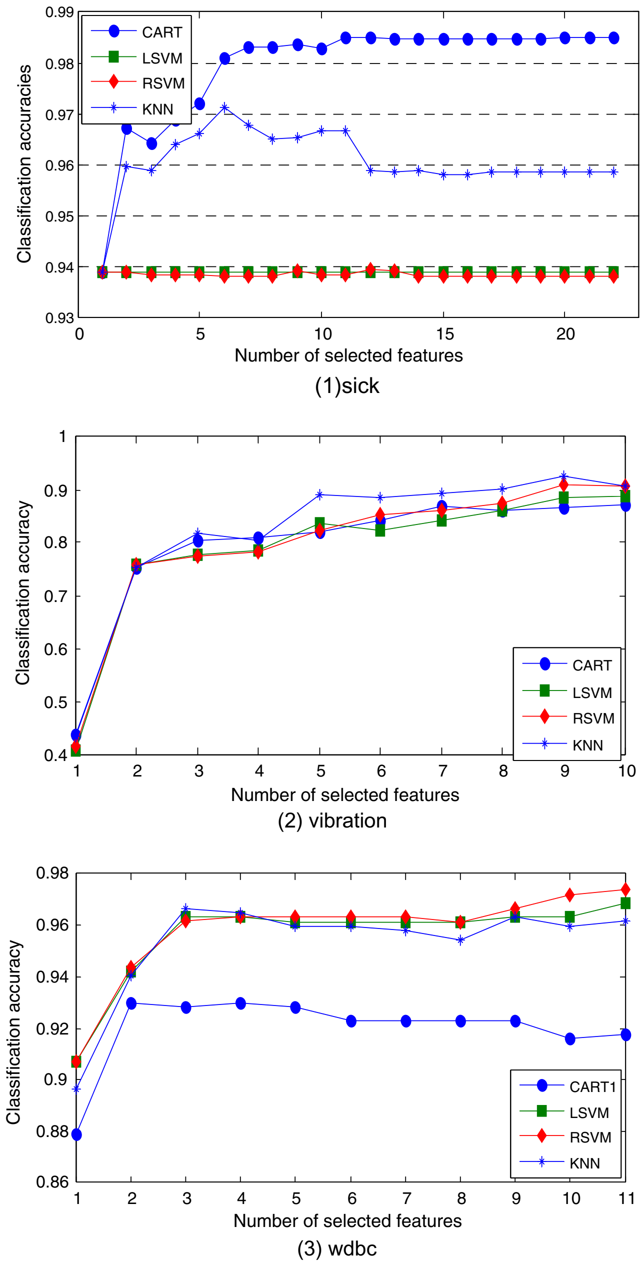


Fig. 6. Variation of classification accuracies with number of selected features.

For sick dataset, the greatest classification accuracies do not occur to the final subset of features, CART arrives at the peak accuracy when 11 feature are selected; KNN reach the peak when 6 features are selected, whereas the accuracies produced by LSVM and RSVM do not vary from the beginning. For vibration, the accuracies increase when the selected features get more and more. It shows all the selected features are useful for classification learning. There is also a peak for data wdbc; 3 features are enough for LSVM, RSVM and KNN. However, 11 features are selected. This fact of overfitting (too many selected features lead to reduction of performance) was once reported in Raudys and Jain (1991). In Peng et al. (2005), Peng et al. introduced a two-stage feature selection algorithm. In the first stage, they employed mRMR to rank the candidate features, and then used a specific classifier to compute accuracy of each subset in the second stage. The subset of features producing the highest accuracy is finally selected.

5.3. Performance comparison of MD, mRMR, mRMD and others

Given a measure of attribute quality, there are a lot of techniques to search the best features with respect to this measure.

We show four related methods in Section 4. We know the computational complexities for these algorithms are different. And we have compared NMI and MD based algorithm with NRS, CFS, FCBF and consistency. In the following, we discuss the performance of NMI integrated with MD, mRMR and mRMR.

Tables 6–9 give the number of selected features and classification performance of NMI integrated with MD, mRMR and mRMR, CFS, FCBF and mRMR. 8 data sets are chosen for these experiments. As computational complexities of NMI_mRMR, NMI_mRMD and MI_mRMR are very high; the databases with small sizes are used here.

Considering classification accuracy, NMI_mRMR and MI_mRMR are better than other algorithms with respect to the four classification algorithms. It shows the strategy of minimal redundancy and

Table 6
Number and accuracy (%) of features selected with different algorithms (CART).

Data set	NMI_mRMR		NMI _MD		NMI_mRMD		CFS		FCBF		MI_mRMR	
	n	Accuracy	*n*	Accuracy	*n*	Accuracy	*n*	Accuracy	*n*	Accuracy	*n*	Accuracy
Heart	3	85.2 ± 6.3	3	85.2 ± 6.3	3	85.2 ± 6.3	6	77.4 ± 7.1	7	77.0 ± 6.9	3	85.2 ± 6.3
Hepatitis	9	93.0 ± 7.1	2	90.8 ± 4.8	2	90.8 ± 4.8	7	93.0 ± 7.1	7	93.0 ± 7.1	7	92.3 ± 6.7
Horse	22	95.9 ± 2.3	2	91.8 ± 3.9	9	96.5 ± 1.8	8	95.9 ± 1.9	8	95.9 ± 1.9	3	96.5 ± 1.3
Iono	19	91.7 ± 5.1	11	89.3 ± 6.6	9	90.6 ± 7.2	4	87.8 ± 7.0	14	88.7 ± 7.1	7	91.2 ± 2.6
Sonar	7	77.9 ± 5.7	3	76.4 ± 7.0	3	76.4 ± 7.0	10	70.6 ± 12.1	19	70.7 ± 14.1	8	77.9 ± 7.5
Wdbc	5	94.0 ± 3.4	2	93.0 ± 2.6	5	93.5 ± 3.2	7	94.0 ± 4.6	11	92.8 ± 4.8	8	94.7 ± 4.3
Wine	5	91.5 ± 4.8	3	91.5 ± 4.8	5	92.0 ± 6.3	10	90.4 ± 6.5	11	89.9 ± 6.3	4	91.5 ± 4.8
Zoo	10	91.8 ± 9.6	4	92.8 ± 9.9	4	92.8 ± 9.7	6	87.8 ± 10.6	9	93.8 ± 10.1	4	90.8 ± 9.1
Average	10	90.1	4	88.9	5	89.7	7	87.1	11	87.7	6	90.0

Table 7
Number and accuracy (%) of features selected with different algorithms (LSVM).

Data set	NMI_mRMR		NMI_MD		NMI_mRMD		CFS		FCBF		MI_mRMR	
	n	Accuracy	*n*	Accuracy	*n*	Accuracy	*n*	Accuracy	*n*	Accuracy	*n*	Accuracy
Heart	6	84.8 ± 6.4	4	83.3 ± 6.4	4	82.2 ± 7.8	7	84.8 ± 5.9	6	82.2 ± 5.5	6	84.4 ± 6.7
Hepatitis	5	88.0 ± 6.1	5	84.5 ± 4.4	5	88.2 ± 5.5	7	90.2 ± 6.6	7	90.2 ± 6.6	7	91.7 ± 8.2
Horse	21	93.0 ± 4.4	2	90.2 ± 4.1	2	90.2 ± 4.1	8	91.0 ± 5.0	8	91.0 ± 5.0	4	93.8 ± 4.7
Iono	29	88.5 ± 6.5	11	85.6 ± 6.6	15	89.8 ± 4.7	14	86.4 ± 5.3	4	83.2 ± 6.4	14	89.8 ± 5.2
Sonar	8	80.3 ± 7.7	7	72.6 ± 7.0	56	80.3 ± 8.7	19	78.4 ± 5.6	10	77.9 ± 7.1	20	87.9 ± 10.5
Wdbc	17	98.3 ± 1.8	19	97.0 ± 1.4	4	96.7 ± 1.9	11	96.3 ± 1.9	7	95.8 ± 2.8	13	97.7 ± 2.2
Wine	13	98.9 ± 2.3	5	98.3 ± 2.7	8	97.7 ± 3.0	11	98.9 ± 2.3	10	98.9 ± 2.3	9	99.4 ± 1.8
Zoo	12	95.4 ± 8.4	4	88.5 ± 12.2	6	93.4 ± 9.5	9	93.4 ± 8.2	6	93.4 ± 8.3	5	93.4 ± 8.2
Average	14	90.9	7	87.5	13	89.8	11	89.9	7	89.1	10	92.3

Table 8
Number and accuracy (%) of features selected with different algorithms (RBFSVM).

Data set	NMI_mRMR		NMI _MD		NMI_mRMD		CFS		FCBF		MI_mRMR	
	n	Accuracy	*n*	Accuracy	*n*	Accuracy	*n*	Accuracy	*n*	Accuracy	*n*	Accuracy
Heart	4	85.9 ± 6.2	3	85.6 ± 6.2	4	85.9 ± 6.2	7	80.7 ± 6.7	6	80.7 ± 5.5	3	85.6 ± 6.2
Hepatitis	5	88.8 ± 5.7	4	89.0 ± 7.0	5	92.2 ± 6.9	7	89.7 ± 5.5	7	89.7 ± 5.5	6	88.8 ± 6.5
Horse	3	92.1 ± 4.8	2	91.8 ± 3.9	2	91.8 ± 3.9	8	91.6 ± 5.1	8	91.6 ± 5.1	5	92.1 ± 4.8
Iono	17	95.2 ± 4.3	11	95.2 ± 4.3	15	94.9 ± 3.9	14	95.2 ± 4.4	4	89.5 ± 3.9	23	96.0 ± 3.4
Sonar	57	88.0 ± 6.8	5	79.8 ± 6.3	42	87.0 ± 7.8	19	79.8 ± 6.0	10	80.3 ± 8.4	48	88.9 ± 5.7
Wdbc	17	98.1 ± 2.3	19	97.9 ± 2.2	4	96.8 ± 2.0	11	96.8 ± 1.8	7	96.5 ± 2.7	15	97.9 ± 2.5
Wine	10	98.9 ± 2.3	5	97.2 ± 3.0	6	98.3 ± 2.8	11	98.9 ± 2.3	10	98.9 ± 2.3	10	98.9 ± 2.3
Zoo	4	94.5 ± 8.3	4	92.4 ± 9.2	4	92.4 ± 9.2	9	95.5 ± 8.3	6	94.5 ± 8.3	4	94.5 ± 8.3
Average	15	92.7	7	91.1	10	92.4	11	91.0	7	90.2	14	92.8

Table 9
Number and accuracy (%) of features selected with different algorithms (KNN).

Data set	NMI_mRMR		NMI _MD		NMI_mRMD		CFS		FCBF		MI_mRMR	
	n	Accuracy	*n*	Accuracy	*n*	Accuracy	*n*	Accuracy	*n*	Accuracy	*n*	Accuracy
Heart	9	83.3 ± 8.2	7	83.3 ± 9.4	6	85.6 ± 8.3	7	83.0 ± 4.7	6	83.3 ± 5.9	10	84.1 ± 7.6
Hepatitis	1	90.2 ± 7.3	2	92.5 ± 6.8	2	92.5 ± 6.8	7	87.5 ± 7.5	7	87.5 ± 7.5	1	90.2 ± 7.3
Horse	10	93.2 ± 3.4	2	90.2 ± 6.1	2	90.2 ± 6.1	8	91.6 ± 4.9	8	91.6 ± 4.9	1	96.2 ± 3.3
Iono	4	92.1 ± 4.9	2	91.2 ± 5.0	2	91.2 ± 5.0	14	86.1 ± 7.6	4	88.7 ± 4.9	3	90.6 ± 4.8
Sonar	29	86.1 ± 9.4	7	81.7 ± 5.9	6	83.1 ± 7.6	19	81.7 ± 5.4	10	81.2 ± 7.6	16	83.1 ± 5.3
Wdbc	23	97.2 ± 1.9	19	96.8 ± 2.0	3	96.7 ± 2.2	11	96.8 ± 2.3	7	95.8 ± 2.0	4	96.8 ± 2.7
Wine	6	97.7 ± 3.0	5	98.3 ± 2.7	5	98.2 ± 4.1	11	97.2 ± 3.0	10	96.6 ± 2.9	6	97.7 ± 3.0
Zoo	3	88.4 ± 9.4	5	86.0 ± 7.3	6	88.3 ± 6.2	9	89.3 ± 7.6	6	88.3 ± 8.2	3	88.4 ± 9.4
Average	11	91.0	6	90.0	4	90.7	11	89.2	7	89.1	6	90.9

maximal relevance is effective for feature subset selection except the high computational complexity. If there is limit in time complexity, NMI_mRMR is preferred.

However, in most cases, computational complexity is very important in machine learning and data mining. NMI_MD can become a substitute for NMI_mRMR as the performance does not reduce too much, but time complexity reduce from $O(N^3)$ to $O(N^2)$. We should also note that the number of the features selected by mRMR is much more than MD. NMI_MD just selects 4 features, while NMI_mRMR selects 10 features for CART algorithm.

Neighborhood mutual information based algorithms are better than CFS anf FCBF. Among 8 data sets, NMI_mRMR get 5 better results than CFS in terms of CART, and the performance of these two algorithms are of difference on other 3 data sets. The similar cases occur to FCBF. In addition, we also perform *t*-test on the experimental results. *t*-test shows that at the 0.1 significance level, the average accuracies derived from NMI mRMR are better than the ones produced with CFS and FCBF in terms of CART and KNN, and no significant difference is observed from the accuracies derived from NMI mRMR and MI mRMR.

In summary, the features selected by NMI_mRMR and MI_mRMR produce the same classification performance, which shows neighborhood mutual information has the same power of feature evaluation as mutual information. mRMR strategy if very useful for feature selection except its high computational complexity. Considering the complexity, maximal dependency can also become a substitute of mRMR.

In gene expression based cancer recognition, datasets usually contain thousands of features and tens of samples. High dimensionality is considered as the main challenge in this domain. We collect several cancer recognition tasks, including DLBCL (a dataset recording 88 measurements of diffuse large B-cell lymphoma described with 4026 array elements), Leukemial1 (a collection of 72 expression measurements with 7129 probes) and SRBCT (the small round blue cell tumors with 88 samples and 2308 attributes). Based on KNN, the recognition rates of these tasks are 94.0%, 77.4% and 58.5%, respectively. Then we perform feature selection on these tasks. For DLBCL, NMI mRMR, CFS and FCBF select 10, 357 and 242 features, respectively. The corresponding recognition rates are 99.0%, 99.0% and 98.3% if KNN is used as the classifier. For Leukemial1, NMI mRMR, CFS and FCBF select 16, 102 and 53 features, and the derived accuracies are 98.6%, 97.5% and 96.1%. Finally, for SRBCT, these algorithms return 14, 70 and 50 features and accuracies are 82.2%, 80.5% and 75.3%, respectively. Comparing the three algorithms, we can get that NMI mRMR selects much less features and yield better recognition rates than CFS and FCBF. The recognition performance after dimensionality reduction is significantly improved. The experimental results show NMI mRMR is effective in dealing with gene recognition tasks.

6. Conclusion and future work

Measures for computing the relevance between features play an important role in discretization, feature selection, decision tree construction. A number of measures were developed. Given its effectiveness, mutual information is widely used and discussed for effectiveness. However, it is difficult to compute relevance between numerical features based on mutual information. In this work, we generalize Shannon's information entropy to neighborhood information entropy and propose the concept of neighborhood mutual information (NMI), which can be directly used to compute relevance between numerical features. We show that the new measure is a natural extension of classical mutual information, thus the new measure can also compute the relevance between discrete variables.

We combine the proposed measure with four classes of strategies for feature subset selection. The computational complexities of these algorithms are also presented. Through extensive experiments, it is shown that the neighborhood mutual information produces the nearly same results as those obtained when applying the classical mutual information. This result shows the significance of numerical variables estimated by discretization and mutual information can also be computed with neighborhood mutual information without discretization. The experimental results exhibit the stability of neighborhood mutual information. Thus neighborhood mutual information is an effective and stable measure for computing relevance between continuous or discrete features. Moreover, we also tested the proposed feature selection algorithms based on NMI. The results show that the features selected with NMI based algorithms are better than those selected with CFS, consistency and FCBF in terms of classification accuracies.

NMI is able to compute the relevance between continuous features and discrete features. Thus it can also be used to compute the significance of features and select features for regression analysis. This forms an interesting topic for further studies.

Acknowledgement

This work is supported by National Natural Science Foundation of China under Grant 60703013 and 10978011, and The Hong Kong Polytechnic University (G-YX3B).

References

Battiti, R. (1994). Using mutual information for Selecting features in supervised neural net learning. *IEEE Transactions on Neural Networks, 5*, 537–550.

Bell, D., & Wang, H. (2000). A formalism for relevance and its application in feature subset selection. *Machine Learning, 41*(2), 175–195.

Blake, C. L., & Merz, C. J. (1998). UCI repository of machine learning databases. <http://www.ics.uci.edu/~mlearn/MLRepository.html>.

Breiman, L. et al. (1993). *Classification and regression trees*. Boca Raton: Chapman and Hall.

Dash, M., & Liu, H. (2003). Consistency-based search in feature selection. *Artificial Intelligence, 151*(1–2), 155–176.

Düntsch, I., & Gediga, G. (1997). Statistical evaluation of rough set dependency analysis. *International Journal of Human Computer Studies, 46*(5), 589–604.

Fayyad, U., & Irani, K. (1993). Multi-interval discretization of continuous-valued attributes for classification learning. In *Proceedings of 13th international joint conference on artificial intelligence. San Mateo, CA: Morgan Kaufmann* (pp. 1022–1027).

Fayyad, U. M., & Irani, K. B. (1992). On the handling of continuous-valued attributes in decision tree generation. *Machine Learning, 8*, 87–102.

Fleuret, F. (2004). Fast binary feature selection with conditional mutual information. *Journal of Machine Learning Research, 5*, 1531–1555.

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research, 3*, 1157–1182.

Hall, M. A. (1999). Correlation-based feature subset selection for machine learning, Ph. D. dissertation, Univ. Waikato, Waikato, New Zealand.

Hall, M. A. (2000). Correlation-based feature selection for discrete and numeric class machine learning. In *Proceedings 17th international conference on machine learning* (pp. 359–366).

Huang, J. J., Cai, Y. Z., & Xu, X. M. (2008). A parameterless feature ranking algorithm based on MI. *Neurocomputing, 71*(1-2), 1656–1668.

Hu, X. H., & Cercone, N. (1995). Learning in relational databases: A rough set approach. *Computational Intelligence, 12*(2), 323–338.

Hu, Q. H., Xie, Z. X., & Yu, D. R. (2007). Hybrid attribute reduction based on a novel fuzzy-rough model and information granulation. *Pattern Recognition, 40*(12), 3509–3521.

Hu, Q. H., Yu, D. R., Liu, J. F., & Wu, C. (2008). Neighborhood rough set based heterogeneous feature subset selection. *Information Sciences, 178*(18), 3577–3594.

Hu, Q. H., Yu, D. R., & Xie, Z. X. (2006). Information-preserving hybrid data reduction based on fuzzy-rough techniques. *Pattern Recognition Letters, 27*(5), 414–423.

Hu, Q. H., Yu, D. R., Xie, Z. X., & Li, X. D. (2007). EROS: Ensemble rough subspaces. *Pattern Recognition, 40*(12), 3728–3739.

Hu, Q. H., Yu, D. R., Xie, Z. X., & Liu, J. F. (2006). Fuzzy probabilistic approximation spaces and their information measures. *IEEE Transactions on Fuzzy Systems, 14*(2), 191–201.

Kalousis, A., Prados, J., & Hilario, M. (2007). Stability of feature selection algorithms: A study on high-dimensional spaces. *Knowledge and Information Systems, 12*(1), 95–116.

Kwak, N., & Choi, C.-H. (2002). Input feature selection for classification problems. *IEEE Transactions on Neural Networks, 13*(1), 143–159.

Kwak, Nojun, & Choi, Chong-Ho (2002). Input feature selection by mutual information based on Parzen window. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 24*(12), 1667–1671.

Liu, H., Hussain, F., & Dash, M. (2002). Discretization: An enabling technique. *Data Mining and Knowledge Discovery, 6*, 393–423.

Liu, X. X., Krishnan, A., & Mondry, A. (2005). An entropy-based gene selection method for cancer classification using microarray data. *BMC Bioinformatics, 6*, 76. doi:10.1186/1471-2105-6-76.

Liu, H., & Setiono, R. (1997). Feature selection via discretization of numeric attributes. *IEEE Transactions on Knowledge and Data Engineering, 9*(4), 642–645.

Liu, H., & Yu, L. (2005). Toward integrating feature selection algorithms for classification and clustering. *IEEE Transactions on Knowledge and Data Engineering, 17*(4), 491–502.

Pawlak, Z. (1991). *Rough sets: Theoretical aspects of reasoning about data*. Dordrecht: Kluwer Academic Publishers.

Pawlak, Z., & Rauszer, C. (1985). Dependency of attributes in information systems. *Bulletin of the Polish Academy of Sciences Mathematics, 33*, 551–559.

Peng, H. C., Long, F. H., & Ding, Chris (2005). Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 27*(8), 1226–1238.

Qian, Y. H., Liang, J. Y., & Dang, C. Y. (2009). Knowledge structure, knowledge granulation and knowledge distance in a knowledge base. *International Journal of Approximate Reasoning, 50*, 174–188.

Qu, G., Hariri, S., & Yousif, M. (2005). A new dependency and correlation analysis for features. *IEEE Transactions on Knowledge and Data Engineering, 17*(9), 1199–1207.

Quinlan, J. R. (1986). *Induction of decision trees, 1*(1), 81–106.

Quinlan, J. R. (1993). *C4. 5: Programming for machine learning*. Morgan Kauffmann.

Raudys, S. J., & Jain, A. K. (1991). Small sample size effects in statistical pattern recognition: recommendations for practitioners. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 13*(3), 252–264.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal, 27*, 379–423. 623–656.

Sikonja, M. R., & Kononenko, I. (2003). Theoretical and empirical analysis of ReliefF and RReliefF. *Machine Learning, 53*, 23–69.

Wang, H. (2006). Nearest neighbors by neighborhood counting. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 28*(6), 942–953.

Wang, H., Bell, D., & Murtagh, F. (1999). Axiomatic approach to feature subset selection based on relevance. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 21*(3), 271–277.

Wettschereck, D., Aha, D. W., & Mohri, T. (1997). review and comparative evaluation of feature weighting methods for lazy learning algorithms. *Artificial Intelligence Review, 11*(1-5), 273–314.

Yao, Y. Y., & Zhao, Y. (2008). Attribute reduction in decision-theoretic rough set models. *Information Sciences, 178*(17), 3356–33731.

Yu, L., & Liu, H. (2004). Efficient feature selection via analysis of relevance and redundancy. *Journal of Machine Learning Research, 5*(Oct), 1205–1224.

Yu, D. R., Hu, Q. H., & Wu, C. X. (2007). Uncertainty measures for fuzzy relations and their applications. *Apply Soft Computing, 7*, 1135–1143.