An Adaptive Neighborhood Partition Full Conditional Mutual Information Maximization Method for Feature Selection

Gaoshuai WANG\textsuperscript{a} (gaoshuai.wang@utbm.fr), Fabrice Lauri\textsuperscript{b} (fabrice.lauri@utbm.fr), Pu WANG\textsuperscript{c} (MagicBeanou@outlook.com), Hongyuan LUO\textsuperscript{d} (hongyuan.luo@ccnu.edu.cn), Amir HAJJAM EL HASSANI\textsuperscript{a} (amir.hajjam-el-hassani@utbm.fr)

\textsuperscript{a} Nanomédicine, Imagerie, Therapeutique - EA4662, Université Bourgogne Franche-Comté, UTBM, F-90010 Belfort, France
\textsuperscript{b} Connaissance et Intelligence Artificielle Distribuée (CIAD) - UMR 7533, Université Bourgogne Franche-Comté, UTBM, F-90010 Belfort, France
\textsuperscript{c} Xi’an XD Switchgear Electric Co.LTD, 710077, Xi’an, China
\textsuperscript{d} School of Information Management, Central China Normal University, Wuhan 430079, China

Corresponding Author:
Gaoshuai WANG
Nanomédicine, Imagerie, Therapeutique - EA4662, Université Bourgogne Franche-Comté, UTBM, F-90010 Belfort, France
Email: gaoshuai.wang@utbm.fr
An Adaptive Neighborhood Partition Full Conditional Mutual Information Maximization Method for Feature Selection

Gaoshuai WANG\textsuperscript{a,*}, Fabrice Lauri\textsuperscript{b}, Pu WANG\textsuperscript{c}, Hongyuan LUO\textsuperscript{d}, Amir HAJJAM EL HASSANI\textsuperscript{a}

\textsuperscript{a}Nanomédecine, Imagerie, Therapeutique - EA4662, Université Bourgogne Franche-Comté, UTBM, F-90010 Belfort, France
\textsuperscript{b}Connaissance et Intelligence Artificielle Distribuées (CIAD) - UMR 7533, Université Bourgogne Franche-Comté, UTBM, F-90010 Belfort, France
\textsuperscript{c}Xi’an XD Switchgear Electric Co.LTD, 710077, Xi’an, China
\textsuperscript{d}School of Information Management, Central China Normal University, Wuhan 430079, China

Abstract

Feature selection is used to eliminate redundant features and keep relevant features, it can enhance machine learning algorithm’s performance and accelerate computing speed. In various methods, mutual information has attracted increasingly more attention as it’s an effective criterion to measure variable correlation. However, current works mainly focus on maximizing the feature relevancy with class label and minimizing the feature redundancy within selected features, we reckon that pursuing feature redundancy minimization is reasonable but not necessary because part of so-called redundant features also carries some useful information to promote performance. In terms of mutual information calculation, it may distort the true relationship between two variables without proper neighborhood partition. Traditional methods usually split the continuous variables into several intervals even ignore such influence. We theoretically prove how variable fluctuation negatively influences mutual information calculation. To remove the referred obstacles, for feature selection method, we propose a full conditional mutual information maximization method (FCMIM) which only considers the feature relevancy in two aspects. For obtaining a better partition effect and eliminating the negative influence of attribute fluctuation, we put up an adaptive neighborhood partition algorithm (ANP) with the feedback of mutual information maximization algorithm, the backpropagation process helps search for a proper neighborhood partition parameter. We compare our method with several mutual information methods on 17 benchmark datasets. Results of FCMIM are better than other methods based on different classifiers. Results show that ANP indeed promotes nearly all the mutual information methods’ performance.

Keywords: Mutual information, feature selection, feature relevancy, neighborhood partition

1. Introduction

Processing the high-dimensional dataset is a vital work in recent years as the rapid growth demand from genetic engineering, image process, pattern recognition, etc. The core idea of dealing with these datasets is to reduce their dimensionality scale by using various techniques (Zhang et al., 2019). At present, dimensionality reduction has two popular ways, the first one is feature extraction, another named feature selection. Feature extraction will transfer data from a high-dimensional space into a low-dimensional space without conserving the original information. The representative algorithms are PCA and LDA (Ayesha et al., 2020; Wang et al., 2021). However, these kinds of methods are not suitable for interpreting the relationship between features and class label. Feature selection picks up some considered important features from the original dataset. About the selection strategies, feature selection techniques are broadly divided into three categories: filter methods, wrapper methods, and embedded methods. Filter methods

\textsuperscript{*}Corresponding author.
Email addresses: gaoshuai.wang@utbm.fr (Gaoshuai WANG), fabrice.lauri@utbm.fr (Fabrice Lauri), MagicBeandou@outlook.com (Pu WANG), hongyuan.luo@ccnu.edu.cn (Hongyuan LUO), amir.hajjam-el-hassani@utbm.fr (Amir HAJJAM EL HASSANI)
use different metrics to assess a feature’s importance and select the high-ranked value features (Kamalov & Thabtah, 2017). Generally, correlation, Gini, and mutual information are the most popular metrics in recent research works. With the existence of highly correlated features, classifiers can run faster and get a more accurate result. Wrapper methods are a series of methods aiming at searching the optimal feature subset via testing all the possible subsets with a specific machine learning method (Thabtah et al., 2020). It consists of two components, 1) selecting a feature subset, 2) using the fixed classification method to assess this feature subset. Due to the complex interaction among features and lacking effective tools to mine features’ relationships, the total test combination number is \(2^n - 1\) when the feature amount is \(n\). With current computation ability, the time consumption is unacceptable for wrapper methods to obtain the optimal result on a high-dimensional dataset. The classification task with wrapper methods could be restricted by time consumption. Hence, researchers are trying to look for more efficient ways to estimate the ideal feature subset (Hoque et al., 2014; Thabtah & Abdelhamid, 2016). Embedded methods are proposed and incorporate identifying the optimal subset as part of training with a classifier. For example, many feature factors trend to a small value after the training process.

After comparison, we consider that the filter methods are more effective for selecting the key features with corresponding criteria. Especially, mutual information based on the information theory is easier to grasp indispensable features. In this paper, we focus on finding a more proper way to calculate the mutual information and maximizing the mutual information between the candidate features and class label.

Based on the change of Shannon entropy (Shannon, 1948), researchers make use of this metric to assess the correlation between feature and class label and name it as mutual information. The mutual information is calculated by Eq. 1:

\[
I(x; y) = H(x) - H(x|y)
\]  
(1)

where \(H(x)\) is the information entropy of \(x\), \(H(x|y)\) is the conditional entropy of \(x\) with \(y\), it represents how much the uncertainty is left when \(y\) is given. The calculation method of conditional entropy is similar to information entropy. Considering the complex interaction in features, conditional mutual information is introduced to the mutual information. Apparently, conditional mutual information can be explained as the mutual information between \(x\) and \(y\) under the existence of \(z\). It’s given in Eq. 2:

\[
I(x; y|z) = H(x|z) + H(y|z) - H(x, y|z)
\]  
(2)

In addition, joint mutual information is also an important concept. It represents the mutual information between \((x, y)\) and \(z\):

\[
I(x, y; z) = I(x; z) + I(y; z|x)
\]  
(3)

Feature selection by mutual information tries to search a subset feature from original features which can own the maximal mutual information \(I(S; y)\), where \(S\) is the set of the selected features. To solve the NP-hard problem, researchers apply the sum of each selected feature’s mutual information with the label to represent \(I(S; y)\). The traditional mutual-information-based method considers three aspects: maximizing the feature relevancy with class label, minimizing the feature redundancy within selected features, and maximizing the conditional feature redundancy between the candidate feature and the selected features. The calculation formation is shown in Eq. 4:

\[
J(x_k) = I(x_k; y) - \beta \sum_{x_j \in S} I(x_j; x_k) + \lambda \sum_{y \in S} I(x_j; y)
\]  
(4)

where \(x_k\) is a candidate feature, \(x_j\) is one of the selected features, \(y\) is the label. \(\beta\) and \(\lambda\) are the non-negative parameters of different items. Based on the popular mutual information framework, in Fig. 1 we can observe some drawbacks in this framework. We categorize the dataset into two types, features are independent with each other, or not independent. When the dataset is the last type, the interaction between two features is so weak, their relationships with labels are like the right side in Fig. 1. This situation makes using mutual information to select the feature turn into a simple task. Once the features are correlated with each other, it’s like the left side. In Eq. 4, it demonstrates that an ideal candidate feature should not be correlated with the selected features, but when the class label exists. However, outside the class label’s information entropy, the mutual information between candidate feature and the selected features has no contribution to the results. Furthermore, even the candidate feature correlated with the selected features, it still contains lots of class label’s information. In our opinion, as long as the feature can provide enough information to
eliminate the uncertainty of class label, it should be considered selected. Therefore, we reckon that the framework in Eq. 4 is improper for selecting the targeted candidate feature.

Fig. 1. Relationships among three features

Apart from designing a more suitable mutual information calculation method, eliminating attribute fluctuation is an important task, too. Present methods usually ignore the negative impact of feature attribute fluctuation on the mutual information calculation. Fig. 2 and Table 1 respectively demonstrates this phenomenon in different views. Part of researchers tries to use the neighborhood mutual information to solve the problem [Hu et al., 2011; Mariello & Battiti, 2018]. Essentially, neighborhood mutual information is the same as mutual information with neighborhood attribute partition. However, it’s not easy for choosing a proper partition distance.

![Fig. 2. Artificial attribute example. There are two probability density curves with different labels within one feature. Ideally, the two curves should not intersect with each other. When some attributes with the same value appear in the intersection distribution range, this feature’s mutual information with class label will decrease. If the fluctuation makes these attributes’ values different, mutual information doesn’t change.](image)

| $f_1$ | $f_2$ | $f_3$ | $y$ |
|-------|-------|-------|-----|
| 1     | 1     | 2     | 0   |
| 2     | 2     | 2     | 0   |
| 4     | 4     | 4     | 0   |
| 5     | 4     | 4     | 1   |
| 7     | 7     | 8     | 1   |
| 8     | 8     | 8     | 1   |

Table 1: Artificial feature example. $I(f_1, y) = 1.0, I(f_2, y) = 0.67, I(f_3, y) = 0.67. f_1, f_2, and f_3 display the influence of attribute fluctuation on mutual information. If 4 and 5 in $f_1$ belong to the same label, mutual information will decrease ($f_1 \rightarrow f_2$). However, if the attributes are still in the same label after the neighborhood partition ($f_2 \rightarrow f_3$), there is no influence on the feature’s mutual information.

To make up for the shortcomings of current research, we propose an adaptive neighborhood partition full conditional mutual information maximization method. The main contributions in this paper are as follows:

1) We propose a novel mutual information method that mainly takes into account the direct relevancy and conditional relevancy between the candidate feature and class label.

2) Based on the neighborhood mutual information theory, the attribute partition method replaces the complex neighborhood mutual information computing process, the definition of the neighborhood is redefined.

3) We propose an adaptive neighborhood partition algorithm with the feedback of MIM method, which can automatically discover a proper attribute partition.

4) Lots of experiments verify the outperformance of FCMIM and ANP.

The structure of the rest part is arranged as follows. Section 2 is the preliminaries of neighborhood rough set and neighborhood mutual information. In Section 3, we review the related works on mutual information and neighborhood partition. Section 4 contains a detailed description of our method. We discuss the experimental results on the 17 benchmark datasets in Section 5. Finally, Section 6 concludes the whole work and puts up the next research plan.
2. Preliminaries

2.1. Neighborhood rough set

Pawlak [1982] first puts up the rough set theory in 1982, which aims at tackling the dataset’s vagueness, imprecision, and uncertainty. However, it’s not suitable for processing continuous datasets due to the probability of samples with the same value being so small. The samples with similar feature values should be categorized into the same class. Based on this idea, Hu et al. [2008] propose the concept of neighborhood rough set.

Define \( U = \{x_1, x_2, ..., x_n\} \) is a set of samples, \( \Delta \) represents a distance function on \( U \), which satisfies \( \Delta(x_i, x_j) \geq 0 \). Given a distance \( \delta \geq 0 \), the neighborhood number of sample \( x_i \) is calculated by \( \delta(x) = \{x_i \mid \Delta(x_i, x_i) \leq \delta\} \). When extended to the two features \( R \) and \( S \), \( \delta_R(x) \) and \( \delta_S(x) \) are the neighborhoods with the certain distance calculation method.

They define the joint neighborhood set between \( R \) and \( S \):

\[
\delta_{(R,S)}(x) = \delta_R(x) \cap \delta_S(x)
\]

In classification work, each sample will be assigned a class label by the decision system. These samples are categorized into \( c_1 \cup c_2 \cup ... \cup c_k = U \), and \( c_i \cap c_j = \emptyset \) if \( i \neq j \). If \( \delta(x) \subseteq c_i \), where \( c_i \) is the subset of samples having the same class label as \( x \), the sample \( x \) is \( \delta \)-neighborhood consistent. And, the \( \delta(x) \) is the lower approximation of \( c_i \). Correspondingly, the upper and lower approximations of class \( c_i \) are written as:

\[
\overline{N}c_i = \{x \mid \delta(x) \cap c_i \neq \emptyset\}, \quad \underline{N}c_i = \{x \mid \delta(x) \subseteq c_i\}
\]

where \( N \) is a neighborhood relation over \( U \). Consequently, the total upper and lower approximations of classification \( C \) are defined as:

\[
\overline{NC} = \bigcup_{i=1}^{k} \overline{N}c_i, \quad \underline{NC} = \bigcup_{i=1}^{k} \underline{N}c_i
\]

Generally, \( \underline{N}c_i \subseteq \overline{N}c_i \), the boundary region of \( c_i \) is defined as:

\[
BN(c_i) = \overline{N}c_i - \underline{N}c_i
\]

If \( BN(c_i) = \emptyset \), it means that all the \( c_i \) samples are classified into \( c_i \), we say \( c_i \) is \( \delta \)-neighborhood consistent.

2.2. Neighborhood mutual information

Based on the neighborhood rough set, Hu et al. [2011] introduce the concept of neighborhood into the information theory. It’s beneficial for applying Shannon’s entropy on the continuous datasets.

Given a set of samples \( U = \{x_1, x_2, ..., x_n\} \) with the continuous or discrete features \( F, S \subseteq F \) is a subset of attributes. Sample \( x_i \)’s neighborhood in \( S \) is defined by \( \delta_S(x_i) \). Afterwards, the neighborhood entropy is defined as:

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

and the average neighborhood entropy can be computed as:

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

If \( \delta = 0 \), \( NH_\delta(S) = H(S) \), where \( H(S) \) is the information entropy. As for two attribute sets \( R, S \subseteq F \), the joint neighborhood of sample \( x_i \) is computed by \( \delta_{R,S}(x_i) \), correspondingly, the joint neighborhood entropy is defined as:

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

Correspondingly, the conditional neighborhood entropy of \( R \) to \( S \) is defined as:

\[
NH_\delta(R|S) = -\frac{1}{n} \sum_{i=1}^{n} \log \frac{||\delta_{R,S}(x_i)||}{n} = NH_\delta(R,S) - NH_\delta(S)
\]
The neighborhood mutual information of $R$ and $S$ is defined as

$$NMI_δ(R; S) = \frac{1}{n} \sum_{i=1}^{n} \log \frac{\|δ_R(x_i)\| \cdot \|δ_S(x_i)\|}{\|δ_{S∪R}(x_i)\|}$$

$$= NH_δ(R) + NH_δ(S) - NH_δ(R, S)$$

$$= NH_δ(R) - NH_δ(R|S) = NH_δ(S) - NH_δ(S|R)$$

By controlling the value of $δ$, we can manipulate the size of neighborhood granularity and further affect the neighborhood entropy and mutual information.

### 3. Related works

#### 3.1. Mutual information

Mutual information as one of the filter methods has attracted researchers’ attention in the past few years. Various works put up their opinions on how to accurately select the key features based on the information theory. Mutual Information Maximization (MIM) method (Lewis, 1992) directly uses the mutual information theory for obtaining the most relevant features with the class label. According to MIM’s view, each feature in the dataset is independent of other features. The specific form is as follows:

$$J(x_k) = I(x_k; y)$$

where $x_k$ is the candidate feature and $y$ represents the class label. Based on the hypothesis of independence in features, MIM just needs to select a few features with the largest mutual information. While most of the features have a certain relation with other features, thus, some methods considering the interaction between features are put up. Battiti (1994) proposes a novel method named Mutual Information Feature Selection (MIFS) which reckons that redundancy has a negative influence on feature selection. The specific form is as follows:

$$J(x_k) = I(x_k; y) - β \sum_{x_j \in S} I(x_j; x_k)$$

where $S$ is the subset of the selected features. $x_j$ is one of the selected features in $S$. The value of $β$ is between 0 to 1. Based on the MIFS, some researchers introduce the optimization method to pursue a better result (Hoque et al., 2014; Kwak & Choi, 2002). These methods share a common problem, it’s hard to find the proper value of $β$. Bad $β$ setting makes $J(x_k)$ is easy close to one of the two terms.

To solve this problem, Peng et al. (2005) use the average redundancy mutual information to represent the redundancy term and name the method as minimal-redundancy-maximal-relevance (mRMR) in 2005. In mRMR, $β$ equals the inverse of the selected feature number.

$$J(x_k) = I(x_k; y) - \frac{1}{|S|} \sum_{x_j \in S} I(x_j; x_k)$$

Different from the above methods, Joint Mutual information (JMI) considers the joint interaction among the candidate feature, selected features, and the class label (Yang & Moody, 1999). Its specific form is defined as:

$$J(x_k) = \sum_{x_j \in S} I(x_j, x_k; y)$$

After simplifying the formula, it’s as follows:

$$J(x_k) = I(x_k; y) - \frac{1}{|S|} \sum_{x_j \in S} I(x_j; x_k) + \frac{1}{|S|} \sum_{x_j \in S} I(x_j; x_k|y)$$

Dynamic Change of Selected Feature with the class (DCSF) (Gao et al., 2018) has a similar formation in Eq. 4, it utilizes a pair of parameters: $β = \frac{1}{|S|}$, $λ = \frac{2}{|S|}$. 

5
Fleuret (2004) proposes the Conditional Mutual Information Maximization (CMIM) method which intends to decrease the redundancy between the candidate feature and the selected features under the class label. Lin & Tang (2006) reckon that the condition mutual information between the candidate feature and the selected feature should be retained. They remove the redundancy factor and use the sum of redundant mutual information. Their method named Conditional Informax Feature Extraction (CIFE) is as follows:

\[
J(x_k) = I(x_k; y) - \sum_{x_j \in S} I(x_j; x_k) + \sum_{x_j \in S} I(x_j; x_k | y) \tag{10}
\]

Gao et al. (2020) put up the Min-Redundancy and Max-Dependency (MRMD), they design the new redundancy term which considers the traditional redundancy term can’t measure the candidate feature accurately, the specific form is as follows:

\[
J(x_k) = I(x_k; y) - \frac{1}{|S|} \sum_{x_j \in S} I(x_j; x_k) + \frac{1}{|S|} \sum_{x_j \in S} I(x_k; y | x_j) \tag{11}
\]

3.2. Neighborhood partition

Besides creating the novel method to select the candidate feature, proper attribute or neighborhood partition plays a vitally important function in feature selection (Sharmin et al., 2019; Tsai & Chen, 2019; Garcia et al., 2012). Neighborhood partition is beneficial for increasing the learning efficiency and classification accuracy (Tahan & Asadi, 2018). Generally, this partition (also called data discretization) is categorized with different criteria, such as static versus dynamic, univariate versus multivariate, supervised versus unsupervised, and splitting versus merging.

Peker & Kubat (2021) use four Chi-square supervised algorithms to partition the continuous datasets on four classifiers. Results show that the performance is better after pre-processing than the original data. Tran et al. (2017) combine the Chi-square method with their potential particle swarm optimization on 10 datasets, the combination can lead to a better performance than applying them in two separate stages.

Wulandari et al. (2019) adopt the merge method to replace the traditional discretization methods for discovering the rare-unusual association rules from a stroke medical dataset. Their experiment indicates that mutual information discretization performs better than the traditional method.

The neighborhood entropy is from the rough set theory which is used to deal with vagueness and uncertainty (Pawlak et al., 1995). Hu & Yu (2009) and Hu et al. (2011) propose the concept of neighborhood entropy and apply the neighborhood mutual information on the discrete and continuous datasets with the fixed partition factor. Xiao et al. (2019) introduce the cuckoo search method to mutual information for discovering the local optimal partition factor. They adopt the symmetrical uncertainty as the loss function to evaluate the partition effect. Sun et al. (2019) propose a novel feature selection method based on neighborhood rough sets using neighborhood entropy-based uncertainty measures for cancer classification.

4. The proposed method

In this section, we propose an adaptive neighborhood partition full conditional mutual information maximization method (ANPPFCMIM) that consists of two parts, one is the novel mutual information term (FCMIM), another is the adaptive neighborhood partition (ANP).

4.1. FCMIM

In previous works, many methods based on mutual information emphasize removing redundancy within the selected features and retaining the relevancy between the selected features and class label. However, a candidate feature can own relevancy and redundancy with the class label and the selected features simultaneously. Relevancy has a more profound influence on prediction, it inspires us to only maximize the relevancy. Generally, \( x_k \) represents the candidate feature, \( x_j \in S \) is one of the already-selected features, \( y \) is the class label. According to the Venn diagram in Fig.1

\[
I(x_k; y) = I(x_k; y | x_j) + I(x_k; x_j; y) \tag{12}
\]
In most cases, \( I(x_k; x_j) \) is regarded as the redundant term and removed. It’s reasonable for excluding \( I(x_k; x_j|y) \) because it has no contribution to offer the information of \( y \). However, when \( I(x_k; x_j) \) is removed, according to Eq. (12), \( I(x_k; x_j; y) \) is also removed, it’s obviously adverse for mutual information calculation.

Feature selection means finding a feature subset in a limited number that contains information as more as possible. Thus, we should focus on discovering the feature which has a large mutual information value with the class label. Redundancy between the candidate feature and the selected features doesn’t demonstrate that the candidate feature is redundant. Our novel method’s term is as follows:

\[
J(x_k) = I(x_k; y) + \frac{1}{|S|} \sum_{x_j \in S} I(x_j; y|x_k)
\]

(14)

where \( I(x_k; y) \) is the primary relevant term, \( \frac{1}{|S|} \sum_{x_j \in S} I(x_j; y|x_k) \) is the secondary relevant term. We present FCMIM selected colored regions in Fig. 3.

Fig. 3. FCMIM’s mutual information calculation regions

Generally, features are not independent, therefore, some mutual information within the selected features repeatedly accumulate and cause a bad influence. We use the average term \( \frac{1}{|S|} \) to weaken its negative influence. Because our method contain the conditional mutual information term and all the selected features are included, we call this method the full conditional mutual information maximization method (FCMIM). When all the features are independent of each other, FCMIM will degenerate to MIM method as the second term in Eq. (14) are the same. We specifically describe the FCMIM method in Algorithm 1.
Algorithm 1: FCMIM

**input**: The training dataset $D$ with original feature set $F = \{f_1, f_2, ..., f_n\}$, class label $Y$ and the required feature number $T$

**output**: The selected subset feature $S_T$

1. $S_T \leftarrow \emptyset$
2. foreach $f_i$ in $F$ do $MI_i = I(f_i; y)$;
3. $f \leftarrow \max( MI)$;
4. $S_T \cup f$
5. $F = F - f$
6. for $i \leftarrow 2$ to $T$ do
   7. for each $f_i$ in $F$ do
   8. $J(f_i) = MI_i + \frac{1}{\|\|} \sum_{f_j \in S} I(f_j; y|f_k)$;
   9. select $f_s$ from $J(f_i)$ with the largest value;
10. $S_T = S_T \cup f_s$
11. $F = F - f_s$
12. return $S_T$

4.2. ANP

In the neighborhood mutual information theory, $\delta$ adjusts the size of the boundary. Using the neighborhood mutual information to evaluate feature’s ability to specify the class label. While the theory doesn’t explain how it promotes the feature selection’s effect. Calculating the number of neighborhoods in a certain distance only changes its proportion. In this part, we will present the details of the adaptive neighborhood partition algorithm. According to the attribute values of datasets, we find that the value of $TOX_{171}$ is large and $COIL_{20}$ is small. We divide the large attribute and multiply the small attribute with the same coefficient. In Fig. 4, each line represents the mutual information between one feature and the class label. With the increment of coefficient, each feature’s mutual information will decrease in $TOX_{171}$ and increase in $COIL_{20}$. For the $TOX_{171}$ dataset, a bigger coefficient means a smaller interval. For the $COIL_{20}$ dataset, it’s the contrary. In the beginning, many features’ mutual information values are very close to each other, after, differences in their values appear. Based on the feature importance assessment mechanism of current mutual information methods, the coefficient tremendously affects feature selection. Therefore, it’s necessary to make an attribute partition before using mutual information theory for selecting the proper features.

![Fig. 4. The influence of attribute partitions on mutual information](image)

**Definition 1.** Given a set of attributes $\{x_1, x_2, ..., x_n\}$ of feature $X$, $n$ is the number of all attributes. $Y = \{y_1, y_2, ..., y_m\}$ in $k$ classes is the class label set, $c_k$ is the attribute number of $k$ th class, $\sum_{i=1}^{k} c_i = n$. The number of attributes which equal to $x_i$ in $X$ is $N_i$, $i \in \{1, 2, ..., n\}$. 

8
Proof 1. As the change happens in terms of \(x_i, x_j \in MI(X; Y)\), we only focus on the change terms.

\[
MI(X; Y)_{x_i} = -p(x_i) \log p(x_i) + p(x_i) \log p(x_i|k) = -\frac{N_i}{n} \log \frac{N_i}{n} + \frac{N_i}{n} \log \frac{N_i}{c_m} = \frac{N_i}{n} \log \frac{n}{c_m},
\]

correspondingly,

\[
MI(X; Y)_{x_j} = \frac{N_j}{n} \log \frac{n}{c_m},
\]

\[
MI(X; Y)_{old} = MI(X; Y)_{x_i} + MI(X; Y)_{x_j} = \frac{N_i}{n} \log \frac{n}{c_m} + \frac{N_j}{n} \log \frac{n}{c_m} = MI(X; Y)_{old}
\]

\[
MI(X; Y)_{new} = MI(X; Y)_{x_i} = \frac{N_i}{n} \log \frac{n}{c_m},
\]

Theorem 2. \(x_i \in m1 th\ class, x_j \in m2 th\ class, m1 \neq m2, m1, m2 \in \{1, 2, ..., k\}, x_i \neq x_j\), if \(x_i \leftarrow x_j\), \(MI(X; Y)_{old} > MI(X; Y)_{new}\)

Proof 2.

\[
MI(X; Y)_{old} = MI(X; Y)_{x_i} + MI(X; Y)_{x_j} = \frac{N_i}{n} \log \frac{n}{N_j} + \frac{N_j}{n} \log \frac{n}{N_j} - \frac{N_i}{n} \log \frac{c_{m1}}{N_i} - \frac{N_j}{n} \log \frac{c_{m2}}{N_j} = \frac{N_i}{n} \log \frac{c_{m1}}{N_i} + \frac{N_j}{n} \log \frac{c_{m2}}{N_j}
\]

\[
MI(X; Y)_{new} = MI(X; Y)_{x_i} = \frac{N_i}{n} \log \frac{n}{N_j} + \frac{N_j}{n} \log \frac{n}{N_j} - \frac{N_i}{n} \log \frac{c_{m1}}{N_i} - \frac{N_j}{n} \log \frac{c_{m2}}{N_j} = \frac{N_i}{n} \log \frac{N_i}{N_j} + \frac{N_j}{n} \log \frac{N_j}{N_j} = MI(X; Y)_{old}
\]

Theorem 1 shows that the attribute partition in the same class label does not influence the mutual information between feature and class label; Theorem 2 demonstrates that partitioning attributes in different class labels into the same interval will decrease the mutual information value. If an attribute appears homogeneously in all classes, there is no function for predicting the class label. Inhomogeneous attribute distribution increases class label’s certainty. The fluctuation of the attribute’s value makes distortion on the inhomogeneous distribution. In neighborhood mutual information theory, the attribute with lots of neighborhoods will have higher mutual information, which magnifies attributes’ importance in the high-density region and makes mutual information can’t change correctly. Besides, as the definition of one attribute’s neighborhood, counting these neighborhoods is a high-computing cost.

Different from the neighborhood mutual information theory, we think neighborhood attribute partition is more effective. The distance \(\delta\) should be replaced by the neighborhood attribute partition parameter \(\theta\). All the neighborhood intervals have the equal width. The attributes within the same interval are the neighborhood with each other, which means they should share the same attribute value.

Assume \(D = \{x_i|i \in N, j \in F\}\) as the dataset, \(x_i\) is the attribute of \(x_i\) in the \(j\) th feature. The \(N\) and \(F\) is the number of total samples and features, respectively. \(\theta\) splits the attributes into \(k\) intervals. We define neighborhood interval set \(N_p\) as fellows:

\[
N_p = \{x_i|i \in I_p\}
\]
where $I_p$ is the $p$ th interval, $p \in \{1, 2, ..., k\}$. All the $x_{ij} \in N_p$ are assigned the new same value, then, it can directly use the traditional information entropy method. In this method, it’s hard to find a suitable interval width, a too-long interval introduces uncertainty from outside on the class label, a too-short interval is not enough to eliminate the uncertainty. It means that there is an optimal interval that can make a classifier get its best result. Inspired by the wrapper methods and backpropagation idea, we propose the adaptive neighborhood partition method. Nearly all the mutual information methods contain MIM’s term, thus, we employ MIM algorithm for selecting the feature subset. KNN classifies the input features and produces its prediction. After, the backpropagation process will adjust the interval partition parameter until training times reach the threshold or the accuracy no longer increases. In the end, we can obtain an optimal $\theta$.

We define the input sample as $x_i$, the predicted label as $\hat{y}_i = f(x_i, \theta)$, the label as $y_i$. The loss and cost function are computed by:

$$
L(y, \hat{y}(\theta)) = \begin{cases} 
1, & y_i \neq \hat{y}_i(\theta) \\
0, & y_i = \hat{y}_i(\theta)
\end{cases}
$$

$$
H(\hat{y}_i(\theta)) = \frac{1}{n} \sum_{i=1}^{n} L(y_i, \hat{y}_i(\theta))
$$

Our goal is:

$$
\hat{\theta} = \arg \min_{\theta} H(\hat{y}_i(\theta))
$$

The process is shown in Algorithm 2:

**Algorithm 2: ANP**

**input**: The initial partition $\theta = 10$, the threshold $k = 50$, learning rate = 0.01, training set $X = \{x_1, x_2, ..., x_n\}$, class label set $Y = \{y_1, y_2, ..., y_n\}$

**output**: $\theta_{opt}$

1. $i \leftarrow 0$, acc $\leftarrow 0$, $\theta_{opt} \leftarrow 0$;
2. while $i < k$ or $\Delta H_{list} > 0.01$ do
3.   $i \leftarrow i + 1$;
4.   Normalization($X$);
5.   $X \leftarrow X \times \theta$;
6.   $f_{set} = \text{MIM}(X)$;
7.   $\text{Acc}_{set} \leftarrow \emptyset$;
8.   $d = 0$;
9.   for $j$ in len($f_{set}$) do
10.      $\text{Model} = \text{KNN}(X_{\text{training}}[\cdot : j], Y_{\text{training}})$;
11.      $\hat{Y} = \text{Model}(X_{\text{validation}})$;
12.      $\text{Acc}_{j} = \text{Acc}_{set} \cup \text{Acc}($ $\hat{Y}(\theta))$;
13.      $\text{Acc}_{set} = \text{mean}($ $\text{Acc}_{set})$;
14.      if $\text{Acc}_{j} > \text{acc}$ then
15.         acc $\leftarrow \text{Acc}_{j}$;
16.         $\theta_{opt} \leftarrow \theta$;
17.         $d = \frac{\Delta \text{Acc}_{set}}{\theta_{opt}}$;
18.      $\theta = \theta + d \times \text{learning rate}$;
19. return $\theta_{opt}$

5. Experimental results and discussion

5.1. Datasets and result analysis

To evaluate our method, we compare our method (FCMIM, ANPFCMIM) with 4 traditional mutual information methods (MIM, CIFE, MIFS, mRMR) and two state-of-art methods (DCSF, MRMD). There are 17 benchmark
datasets from UCI (Dua & Graff, 2019). These datasets are in various types, such as biological data (ALLAML, Colon, GLIOMA, Lung, Lung_discrete, Lymphoma, Leukemia, Prostate_GE and TOX_171), handwriting data (USPS), text data (BASEHOCK, PCMAC, and RELATHE), and other datasets. In these datasets, 7 of them are discrete, the rest is continuous. Besides, the range of attribute value is quite different from each other. To eliminate the influence of attribute value’s difference, in most cases, we normalize the dataset and then utilize the attribute partition parameter \( \theta \) to search the proper neighborhood interval. When the discrete dataset contains a few different values, it’s unnecessary to normalize it. The parameter \( \theta \) relies on the characteristic of the dataset. Searching good \( \theta \) value is a high-computing cost process, for accelerating it, we employ MIM to select the candidate feature, because MIM is not only fast and but also the foundation of the most mutual information methods. To avoid the gradient descending method sometimes can’t convergence due to too much uncertainty in results, we set a threshold \( k \) (50) as the criterion to stop the search process. For the classifiers, we adopt four popular machine learning methods, k-nearest neighborhood (k=3), SVM(kernel='linear'), Random Forest(RF), and Gaussian Bayes(GB). All the training and test works are on a server platform with 64-core CPU and 32 GB memory. Similar to the work of Gao et al. (2018), we employ the first 30 features’ average accuracies and the highest accuracies as the evaluation metrics. The 5-fold cross-validation is used for validation. The framework in this research is presented in Fig. 5 which demonstrates the process of model training. In the feature selection block, we have two sets of experiments, the set in the blue box is to display FCMIM and ANPFCMIM’s performances over other methods. The set in the red box is to verify ANP’s generalization on different methods.

In the beginning, we display the average accuracies and their standard deviations in Tables 3-6. The average accuracy indicates the average value of the 17 datasets within the same method. The paired two-tailed t-test is implemented in this research. The W/T/L represents the win/tie/loss statistics of the proposed method relative to other methods. The bold items in tables mean that they are the best results. The ‘+’ in FCMIM method indicates FCMIM’s performance is better than other methods except ANPFCMIM.

As we can see in Tables 3-6 ANPFCMIM achieves the highest average accuracies on the four classifiers. They are 84.85%, 86.99%, 87.15% and 80.55%, respectively. In Table 3 ANPFCMIM outperforms MIM, MIFS, CIFE, mRMR, DCSF, MRMD, and FCMIM by up to 8.60%, 4.41%, 11.02%, 3.16%, 3.25%, 4.68% and 0.96%. ANPFCMIM obtains 12 times the highest accuracies in the 17 datasets, FCMIM has four times, MIFS owns two times, mRMR gets two times, and DCSF and MRMD each only have one time. Except for ANPFCMIM, FCMIM owns 10 times the highest accuracies.

In Table 4, our method is better than MIM, MIFS, CIFE, mRMR, DCSF, MRMD, and FCMIM by up to 10.1%, 8.59%, 10.93%, 4.92%, 5.57%, 6.89%, and 0.96%. ANPFCMIM performs better than other methods on 13 datasets, FCMIM is 4, MIFS is 1, mRMR is 2, and MRMD is 2, and DCSF is 1. Except for ANPFCMIM, FCMIM owns 12 best performances in all datasets.

In Table 5, our method outperforms MIM, MIFS, CIFE, mRMR, DCSF, MRMD and FCMIM by up to 7.68%, 5.27%, 7.24%, 4.02%, 2.30% 4.06% and 1.05%. ANPFCMIM owns 13 times the best performance, FCMIM gets four times, and the rest of the methods achieve one except MIFS, respectively. Except for ANPFCMIM, FCMIM also achieves 13 best performances.

In Table 6 ANPFCMIM performs better than MIM, MIFS, CIFE, mRMR, DCSF, MRMD, and FCMIM by up to 8.71%, 9.10%, 9.68%, 7.83%, 3.20%, 8.08%, and 1.6%. ANPFCMIM has 15 times the highest average accuracies, FCMIM owns five times, MRMD has one time. Except for ANPFCMIM, FCMIM also achieves 14 times the best accuracies.

In Fig. 6-7 these radar figures represent the win/tie/loss of FCMIM and ANPFCMIM to other methods. We can see that both FCMIM and ANPFCMIM perform better than other methods. And, ANPFCMIM shows its superiority than FCMIM according to the win in Fig. 7 and tie in Fig. 6-7.

In Tables 3-6 there is an obvious phenomenon that one dataset’s highest accuracy method with different classifiers is not always the same. For example, In the KNN method, the best performance of Isolet is MIFS, while, DCSF achieves a better result than other algorithms in SVM. Similarly, mRMR gets the two highest accuracies of PCMAC in KNN and SVM, however, its performances in the GB and RF are not as good as previous classifiers. Consequently, we can conclude that the classifier indeed has a certain influence on the results. Adopting multiple methods to evaluate the algorithm’s ability is necessary.

To clearly observe the process of various methods’ specific influence on the prediction, we select four typical datasets: COIL20, Lung, Isolet, and USPS. In Fig. 8(a)-8(d) the x-axis represents the number of features from 1 to
30, the y-axis is the accuracy. Intuitively, with the increment of feature number, all the methods’ accuracies grow with different speeds. In COIL20, Isolet and USPS, the growth rate slows down as the feature number is more than fifteen. In Lung, Isolet, and USPS, we can see that the proposed method’s plot is higher than other algorithms. In Fig. 8(a), the performance curves of ANPFCMIM and FCMIM methods are quite close to each other. What’s more, all the methods except for MIM have the same tendency. We reckon that these datasets make methods select the same subset features. In Fig. 8(b), the feature’s growth doesn’t bring the rapid increment of accuracy as other datasets because the initial few features have included necessary information.

In Table 7, we summarize the highest accuracy of 17 datasets on four classifiers. Among these results, ANPFCMIM achieves 12 of 17 the best performances and has the highest average accuracy. CIFE, mRMR, DCSF and MRMD have one time, respectively. FCMIM owns four times. Except for ANPFCMIM, FCMIM also has 12 of 17 highest average accuracies.

In Table 7, we summarize the highest accuracy of 17 datasets on four classifiers. Among these results, ANPFCMIM achieves 12 of 17 the best performances and has the highest average accuracy. CIFE, mRMR, DCSF and MRMD have one time, respectively. FCMIM owns four times. Except for ANPFCMIM, FCMIM also has 12 of 17 highest average accuracies.

Otherwise, we present our method’s result comparison between the full features and the selected features in Table 8. In this Table, we choose the highest accuracy within the 30 selected features as our result. From the perspective of the average accuracy, the ANPFCMIM has a better performance than the result produced by the full feature. Primarily, KNN and GB results obtain more promotion than the other two classifiers. For the specific dataset, the selected features can significantly promote its prediction accuracy, which means the redundant features are removed; Sometimes, the accuracy will be a little worse than all features’. Rarely, did the selected features decrease the accuracy largely compared with the full features’ because part of the relevant features could be not included.

5.2. Generalization of ANP

We have verified the effect of ANP in Table 3-6. In this section, we compare ANP with the fixed partition (FP) on all the methods to display its generalization. About FP, we only use one number, which equals 10, to split the attributes. According to Table 2, there are ten continuous datasets in our experiment. In these datasets, their attributes are quite different. We normalize all the continuous datasets before selecting the feature for alleviating the difference in features. Fig. 9 is the heatmap that represents the outperformance of ANP’s accuracy over FP’s, the value in each grid cell equals ANP’s average accuracy - FP’s average accuracy. Intuitively, in most cases, the accuracy is promoted
Table 2: The summary of benchmark datasets

| Number | Name     | Instances | Features | Classes | Data type |
|--------|----------|-----------|----------|---------|-----------|
| 1      | ALLAML   | 72        | 7129     | 2       | continuous|
| 2      | COIL20   | 1440      | 1024     | 10      | continuous|
| 3      | GLIOMA   | 50        | 4434     | 4       | continuous|
| 4      | Isolet   | 1560      | 617      | 26      | continuous|
| 5      | Lung     | 203       | 3312     | 5       | continuous|
| 6      | Madelon  | 2600      | 500      | 2       | continuous|
| 7      | Prostate_GE | 102     | 5966     | 2       | continuous|
| 8      | TOX_171  | 171       | 5748     | 4       | continuous|
| 9      | USPS     | 9298      | 256      | 10      | continuous|
| 10     | WarpPIE10p | 210     | 2420     | 10      | continuous|
| 11     | BASEHOCK | 1993      | 4862     | 2       | discrete  |
| 12     | Colon    | 62        | 2000     | 2       | discrete  |
| 13     | Leukemia | 72        | 7070     | 2       | discrete  |
| 14     | Lung_discrete | 73     | 325      | 7       | discrete  |
| 15     | Lymphoma | 96        | 4026     | 9       | discrete  |
| 16     | PCIMAC   | 1943      | 3289     | 2       | discrete  |
| 17     | RELATHE  | 1427      | 4322     | 2       | discrete  |

Table 3: Average accuracy (mean ± std) on KNN

| Dataset    | MIM     | MIHS    | CIJE    | mRMR    | DCSF    | MRMD    | FCMM    | ANPFCCMM |
|------------|---------|---------|---------|---------|---------|---------|---------|----------|
| ALLAML     | 94.80±5.71 | 96.16±4.22 | 84.14±8.11 | 96.34±5.05 | 96.32±9.66 | 89.44±5.78 | 96.46±3.05 (+) | 97.32±2.67 |
| COIL20     | 77.77±14.45 | 85.75±12.98 | 86.89±12.74 | 89.91±14.2 | 90.54±13.8 | 87.77±14.16 | 91.06±13.85 (+) | 93.43±11.62 |
| GLIOMA     | 74.66±10.52 | 74.22±12.26 | 77.33±12.72 | 82.66±10.13 | 80.44±13.76 | 80.44±13.76 | 83.55±9.85 (+) | 84.05±8.65 |
| Isolet     | 46.5±12.52 | 66.9±21.75 | 46.22±9.47 | 64.39±15.2 | 60.27±13.6 | 61.68±16.99 | 60.44±13.88 | 65.22±13.08 |
| Lung       | 90.72±5.07 | 92.63±4.06 | 82.22±3.32 | 93.63±6.3 | 93.29±3.47 | 90.94±5.18 | 94.23±4.99 (+) | 96.43±7.62 |
| madelon    | 74.74±9.04 | 51.97±1.59 | 61.49±4.03 | 51.65±1.83 | 78.75±9.41 | 53.28±2.35 | 79.88±9.08 (+) | 82.02±9.29 |
| Prostate_GE| 91.72±4.61 | 92.47±4.88 | 84.94±5.48 | 92.76±3.72 | 92.25±4.45 | 92.04±5.32 | 93.22±4.51 | 93.62±4.79 |
| TOX_171    | 62.04±8.08 | 72.50±10.58 | 70.25±8.93 | 73.07±9.18 | 71.28±8.74 | 69.54±13.62 | 73.84±10.00 (+) | 74.93±12.20 |
| USPS       | 59.94±10.79 | 82.16±1.12 | 76.88±13.91 | 80.52±14.21 | 79.38±13.65 | 77.52±17.14 | 81.26±8.41 | 83.26±7.23 |
| warpPIE10P | 78.14±11.87 | 83.75±16.16 | 73.65±10.65 | 89.36±14.44 | 86.23±13.37 | 82.73±19.15 | 85.97±12.81 | 88.65±13.02 |
| BASEHOCK   | 85.88±5.46 | 83.71±7.6 | 75.99±3.01 | 84.38±10.31 | 85.37±8.59 | 86.54±7.99 | 85.74±7.15 | 85.74±7.15 |
| colon      | 79.47±9.93 | 75.43±11.72 | 76.49±9.39 | 78.77±8.21 | 80.35±8.69 | 80.00±10.02 | 82.98±11.34 (+) | 82.98±11.34 |
| leukemia   | 94.85±6.64 | 92.87±6.91 | 89.09±7.39 | 92.15±6.94 | 94.54±4.3 | 92.72±5.06 | 95.15±4.22 (+) | 95.15±4.22 |
| lung_discrete | 68.18±12.08 | 80.60±14.84 | 62.87±12.75 | 79.99±11.89 | 78.48±15.07 | 79.24±14.16 | 83.33±15.54 (+) | 83.33±15.54 |
| lymphoma   | 72.98±13.18 | 87.81±10.30 | 66.66±10.32 | 86.09±8.97 | 85.74±10.93 | 84.71±12.55 | 89.31±10.24 (+) | 89.31±10.24 |
| PCIMAC     | 79.10±5.14 | 79.70±10.15 | 71.96±6.49 | 83.53±5.44 | 79.14±7.66 | 80.07±8.33 | 78.90±5.17 | 78.90±5.17 |
| RELATHE    | 69.07±6.52 | 75.07±6.69 | 66.08±3.17 | 72.45±8.1 | 68.33±4.8 | 74.97±7.36 | 69.91±4.37 | 69.91±4.37 |

Average accuracy: 76.23 ± 80.44 ± 73.83 ± 81.69 ± 81.60 ± 80.17 ± 83.89 (+) ± 84.85

W/T/L(FCMIM): 14/1/2 13/1/3 170/0 11/3/3 10/6/1 13/0/4 13/0/4 9/8/0

W/T/L(ANPFCCMM): 14/1/2 13/3/1 170/0 11/4/2 12/4/1 13/0/4 9/8/0
| Dataset     | MIM | MFNS | CIFE | mRMR | DCSF | MRMD | FCIMIM | ANPFCMIM |
|-------------|-----|------|------|------|------|------|--------|----------|
| ALLAML      | 96.66±3.88 | 95.91±4.13 | 90.75±5.93 | 94.54±5.3 | 92.42±5.91 | 93.78±3.99 | 96.66±3.86(+) | 97.35±2.56 |
| COIL20      | 82.26±14.6 | 89.32±10.92 | 91.39±11.06 | 92.75±12.11 | 93.07±12.53 | 91.13±13.19 | 93.13±12.3(+) | 94.60±11.84 |
| GLIOMA      | 78.03±11.04 | 79.33±9 | 78.66±10.10 | 84.00±10.83 | 82.44±7.98 | 82.44±11.61 | 86.44±8.69(+) | 87.81±6.21 |
| Islet       | 52.71±14.15 | 62.40±16.31 | 58.39±13.12 | 70.17±17.05 | 72.43±17.52 | 70.17±19.35 | 64.81±14.99 | 68.81±10.27 |
| lung        | 91.36±3.93 | 85.46±5.11 | 80.43±5.35 | 88.08±7.55 | 89.12±5.28 | 87.81±6.53 | 90.00±3.00(+) | 94.60±3.48 |
| madelon     | 76.07±9.33 | 51.38±1.61 | 66.90±5.54 | 51.72±1.5 | 81.51±4.9 | 54.09±4.26 | 82.27±7.70(+) | 82.20±6.39 |
| prostate_GE | 91.29±4.34 | 92.04±4.14 | 91.54±9.29 | 91.93±5.18 | 92.58±4.65 | 94.19±4.12 | 95.16±5.05(+) | 95.46±5.61 |
| TOX_71      | 60.15±9.84 | 70.19±6.64 | 73.71±6.74 | 71.47±6.16 | 72.05±7.16 | 73.52±13.88 | 75.12±8.05(+) | 77.40±6.91 |
| USPS        | 64.91±10.78 | 80.85±15.41 | 80.41±12.85 | 83.05±14.25 | 81.29±13.33 | 73.96±16.43 | 84.17±13.93(+) | 85.84±12.33 |
| warpPIE10P  | 82.32±11.1 | 88.62±12.86 | 83.22±16.21 | 89.68±12.25 | 90.37±13.09 | 87.13±15.71 | 88.14±12.78 | 91.14±10.78 |
| BASEHOCK    | 88.05±5.56 | 86.37±5.17 | 83.32±3.34 | 88.4±6.36 | 88.36±5.81 | 88.39±6.99 | 88.58±5.78(+) | 88.58±5.78 |
| colon       | 81.05±7.52 | 85.08±7.45 | 82.98±9.19 | 82.98±9.77 | 85.61±10.61 | 85.96±8.63(+) | 85.96±8.63(+) | 84.96±8.85 |
| leukemia    | 94.39±5.07 | 95.91±5.91 | 97.88±3.05 | 94.84±3.27 | 96.21±3.54 | 95.30±4.47 | 94.69±4.85 | 94.69±4.85 |
| lung_discrete | 71.66±14.78 | 80.15±10.79 | 68.33±13.71 | 80.60±12.31 | 79.09±13.26 | 79.09±11.83 | 85.14±7.46(+) | 85±14.76 |
| lymphoma    | 75.05±12.33 | 86.09±9.49 | 79.19±8.52 | 87.93±8.29 | 88.27±7.78 | 86.09±8.09 | 89.31±7.63(+) | 89.31±7.63 |
| PCMAC       | 83.91±3.52 | 84.70±4.63 | 79.83±3.21 | 84.94±3.88 | 83.52±3.19 | 84.76±5.05 | 83.65±4.01 | 83.65±4.01 |
| RELATHE     | 74.56±3.59 | 78.08±5.45 | 71.57±3.29 | 77.39±5.16 | 75.75±4.21 | 79.40±5.45 | 76.73±3.53 | 76.37±3.53 |
| Average accuracy | 76.81 | 74.80 | 76.06 | 82.07 | 81.42 | 80.10 | 85.94 | 79.1/1 |
| W/T/LFCIMIM | 12/4/1 | 14/1/2 | 160/1 | 13/3/1 | 10/5/2 | 10/4/3 | 10/4/3 | 7/9/1 |
| W/T/LANPFCMIM | 12/4/1 | 14/1/2 | 160/1 | 14/2/1 | 12/3/2 | 10/4/3 | 10/4/3 | 7/9/1 |

Table 5: Average accuracy (mean ± std) on Random Forest
Table 6: Average accuracy (mean ± std) on GB

| Dataset   | MIM     | MIFS    | CIFE    | mRMR    | DCSF    | MRMD    | FCMM    | ANPFCMIM |
|-----------|---------|---------|---------|---------|---------|---------|---------|----------|
| ALLAML    | 96.14±3.78 | 95.16±11.94 | 87.22±17.03 | 96.38±13.17 | 95.15±5.72 | 92.58±15.34 | 97.15±3.66± | 97.35±3.72 |
| COIL20    | 64.54±12.34 | 70.82±8.41 | 73.36±9.61 | 82.50±13.43 | 82.99±12.24 | 76.64±11.77 | 83.15±11.87± | 87.53±10.20 |
| GLIOMA    | 72.00±11.34 | 64.00±13.29 | 68.22±9.38 | 72.22±15.88 | 74.44±10.90 | 71.55±13.65 | 77.44±8.51± | 80.25±7.28 |
| Islet     | 42.38±11.94 | 57.28±15.33 | 53.13±11.23 | 58.11±13.02 | 58.72±13.44 | 55.20±14.90 | 60.42±13.3± | 64.12±9.82 |
| lung      | 90.49±4.09 | 71.36±8.41 | 84.91±4.24 | 92.13±5.26 | 91.53±5.13 | 90.56±5.22 | 94.59±3.84± | 95.10±3.91 |
| madelon   | 59.86±2.89 | 52.10±1.58 | 58.59±3.16 | 52.11±1.63 | 59.37±2.21 | 52.53±1.53 | 60.99±1.65± | 63.58±1.42 |
| Prostate GE | 93.01±3.82 | 69.46±15.17 | 89.03±5.86 | 84.40±10.96 | 92.79±3.97 | 91.50±5.49 | 94.08±3.83± | 95.35±3.89 |
| TOX_1T1   | 65.19±5.90 | 67.50±9.21 | 64.03±6.64 | 67.82±8.95 | 69.35±10.82 | 70.44±12.39 | 70.63±8.67± | 73.63±5.94 |
| USPS      | 54.99±8.58 | 70.56±10.93 | 67.16±9.27 | 72.64±10.34 | 73.29±11.01 | 77.47±12.62 | 73.53±11.04 | 78.53±12.26 |
| warpPIE10P | 64.33±11.54 | 78.14±16.31 | 59.69±10.16 | 81.85±13.51 | 83.54±16.27 | 76.56±14.47 | 84.22±14.9± | 87.94±12.05 |
| BASEHOCK  | 85.49±9.16 | 77.39±6.64 | 81.80±2.52 | 85.98±9.36 | 84.80±9.92 | 86.73±8.45 | 86.26±7.97 | 86.26±7.97 |
| colon     | 80.35±10.34 | 80.52±12.01 | 75.43±12.18 | 82.28±9.56 | 82.45±10.47 | 76.14±11.02 | 84.38±7.63± | 84.38±7.63 |
| leukemia  | 90.60±6.61 | 88.93±4.94 | 90.45±5.67 | 90.60±5.55 | 90.75±6.37 | 89.24±6.26 | 91.05±7.28± | 91.05±7.28 |
| lung_disc | 57.27±7.7 | 58.93±11.35 | 55.60±10.34 | 61.51±13.11 | 63.63±12.86 | 65.45±9.06 | 70.60±16.46 | 70.60±16.46 |
| lymphoma  | 65.63±12.67 | 67.12±15.22 | 66.89±15.33 | 71.6±15.84 | 69.54±16.18 | 68.16±15.08 | 72.06±13.57± | 72.06±13.57 |
| PCMAC     | 74.11±4.31 | 74.08±4.03 | 68.09±5.61 | 73.73±4.77 | 74.92±4.67 | 74.23±3.92 | 75.22±4.68± | 75.22±4.68 |
| RELATHE   | 65.02±4.5 | 71.29±5.88 | 61.24±4.33 | 67.98±5.43 | 67.74±4.25 | 70.50±7.87 | 66.41±3.62 | 66.41±3.62 |

| Average accuracy | 71.84 | 71.45 | 70.87 | 72.84 | 77.35 | 72.47 | 78.95± | 80.55 |
|------------------|-------|-------|-------|-------|-------|-------|-------|-------|
| W/T/L(FCMIM)     | 12/5 | 12/4 | 16/1 | 11/5 | 11/5 | 12/4 |       |       |
| W/T/L(ANPFCMIM)  | 14/3 | 14/2 | 16/1 | 13/3 | 13/3 | 14/2 | 8/3 |       |

Fig. 6. Performance comparison between FCMIM and previous methods. Win/tie/loss indicates FCMIM is better/equal to/worse than another method.
Fig. 7. Performance comparison between ANPFCMIM and previous methods. Win/tie/loss indicates ANPFCMIM is better/equal to/worse than another method.

| Dataset       | MIM  | MIFS | CIFE | mRMR | DCSF | MRMD | FCMIM | ANPFCMIM |
|---------------|------|------|------|------|------|------|-------|----------|
| ALLAML        | 96.66| 95.91| 90.75| 96.38| 95.15| 93.78| 97.13  | 97.35    |
| COIL20        | 82.26| 89.32| 91.39| 92.75| 93.07| 91.13| 93.13  | 94.60    |
| GLIOMA        | 78.00| 81.11| 78.66| 84.00| 82.44| 82.44| 86.44  | 88.36    |
| Isolet        | 52.71| 66.92| 58.49| 70.17| 72.43| 70.17| 64.81  | 69.81    |
| lung          | 91.36| 86.44| 84.91| 89.18| 91.53| 87.81| 94.59  | 95.10    |
| madelon       | 76.07| 52.32| 66.90| 52.74| 81.51| 54.09| 82.27  | 85.27    |
| Prostate_GE   | 93.01| 92.79| 91.50| 93.76| 92.79| 94.19| 95.10  | 95.83    |
| TOX_171       | 66.15| 72.50| 73.71| 73.07| 72.05| 73.52| 75.12  | 77.4     |
| USPS          | 64.89| 83.16| 80.41| 83.05| 81.93| 79.63| 84.17  | 85.64    |
| warpPIE10P    | 83.28| 88.90| 88.04| 90.95| 91.37| 87.77| 90.73  | 91.14    |
| BASEHOCK      | 88.80| 86.39| 83.32| 88.69| 89.15| 88.68| 89.20  | 89.20    |
| colon         | 81.05| 85.08| 82.98| 82.98| 82.98| 85.61| 85.96  | 85.96    |
| leukemia      | 94.85| 97.88| 97.88| 95.15| 96.21| 95.30| 95.15  | 95.15    |
| lung_discrete | 75.15| 80.60| 69.54| 83.93| 84.99| 79.24| 85.50  | 85.50    |
| lymphoma      | 76.43| 87.81| 79.19| 90.11| 89.77| 86.09| 89.31  | 89.31    |
| PCMAC         | 84.20| 84.70| 80.26| 84.99| 83.52| 84.76| 84.94  | 84.94    |
| RELATHE       | 74.56| 78.08| 71.57| 77.39| 75.77| 79.4 | 76.37  | 76.37    |
| average       | 79.97| 82.94| 80.56| 84.08| 85.69| 83.15| 86.25  | 87.41    |

Table 7: The highest average accuracies (%) of four classifiers
Fig. 8. The performance of different methods on four datasets
| Dataset       | ALLAML | BASEHOCK | col | GLEDMA | Faesicl | leukemia | lung | lymphoma | madelon | PCMAC | prostate_CGE | RELATHE | TOX_J71 | USPS | warpPIE10P | average |
|--------------|--------|----------|-----|--------|---------|----------|------|----------|---------|-------|-------------|---------|---------|------|-----------|---------|
| KNN          | 81.79  | 67.64    | 92.57 | 78.81  | 78.00   | 88.57    | 96.52| 101.33   | 56.62   | 67.89 | 81.55       | 75.68   | 61.01   | 97.01| 95.71     | 81.59   |
| FCMIM        | 100.00 | 85.74    | 98.14 | 100.00 | 100.00  | 93.15    | 98.36| 87.81    | 90.11   | 84.56 | 93.22       | 77.62   | 82.26   | 83.75| 96.82     | 89.19 (+)|
| ANPFCIMIM    | 97.14  | 87.25    | 98.40 | 80.71  | 76.00   | 96.30    | 98.65| 89.46    | 80.40   | 86.90 | 91.22       | 80.40   | 82.69   | 83.75| 97.63     | 90.43 (+)|
| SVM          | 100.00 | 94.84    | 98.26 | 88.97  | 100.00  | 95.45    | 99.05| 88.89    | 88.95   | 89.85 | 93.54       | 79.25   | 84.22   | 83.54| 96.82     | 90.36 (+)|
| ANPFCIMIM    | 96.07  | 96.32    | 98.50 | 91.21  | 100.00  | 96.37    | 93.05| 92.93    | 68.94   | 89.45 | 95.45       | 81.25   | 88.85   | 85.73| 99.15     | 91.40 (+)|
| RF           | 100.00 | 93.47    | 99.07 | 80.48  | 93.33   | 94.69    | 93.05| 83.00    | 71.58   | 82.54 | 94.00       | 81.58   | 87.13   | 88.43| 88.90     | 88.59 (+)|
| ANPFCIMIM    | 100.00 | 95.34    | 100.00| 96.94  | 90.42   | 95.67    | 93.00| 86.09    | 79.38   | 92.54 | 98.36       | 84.91   | 88.46   | 88.46| 90.51     | 90.51 (+)|
| GB           | 100.00 | 93.37    | 97.35 | 94.67  | 100.00  | 100.00   | 93.75| 94.69    | 98.36   | 84.54 | 96.77       | 91.00   | 86.53   | 92.32| 93.65     | 95.47 (+)|
| ANPFCIMIM    | 100.00 | 95.34    | 100.00| 89.47  | 94.87   | 98.84    | 98.36| 93.00    | 92.32   | 95.47 | 97.42       | 84.00   | 83.01   | 92.32| 95.47     | 88.94 (+)|
| average      | 90.43  | 89.19 (+)| 90.36| 90.26 (+)| 91.40 (+)| 90.51 (+)| 91.77| 88.59 (+)| 90.36 (+)|

Table 8: The result comparison between all feature input and the selected feature input after applying ANP. Of course, sometimes, ANP’s performance is equal to FP’s, even a little worse. More than half of the promotion is about 0-3%, more than a third is above 3%, sometimes, it’s more than 10%. In addition, ANP’s promotion is more apparent on MIM, MIFS and CIFE. It is because the terrible performance of FP provides ANP with huge promotion space.

5.3. Complexity analysis

Assume the dataset owns \( M \) features and \( N \) instances. The targeted number of the selected features is \( k \). In the practical calculation process, we don’t know the probability density distribution, so the time consumption of mutual information, conditional mutual information, and joint mutual information are different, while time complexity is the same. We define the time complexity of mutual information between two variables as \( O(N) \), correspondingly, MIM’s time complexity can be represented as \( O(MN) \). Similarly, the time complexity of MIFS, CIFE, mRMR, JMI, MRMD, and our method is \( O(kMN) \). Totally, nearly all methods are on the same time complexity level, however, their time consumption are different given the dataset scale in this research. The method which contains conditional mutual information and joint mutual information will cost more time. Thus, in these seven methods, MIM’s time consumption is the best; MIFS, JMI, and mRMR are in the second place; Our method is the third; CIFE and MRMD are in the last place. We reckon that our method’s time complexity and time consumption are acceptable.

6. Conclusions

We propose an adaptive neighborhood partition full conditional mutual information maximization method for the feature selection. Instead of maximizing relevancy and minimizing redundancy, the full conditional mutual information maximization mainly focuses on the relevancy among candidate feature, selected features, and class label. As for processing the continuous data and eliminating the attribute fluctuation’s negative influence on mutual information calculation, we apply the adaptive neighborhood partition with the feedback of MIM method to search for a proper partition parameter. Compared with the neighborhood mutual information, ANP owns a more simple mechanism to partition the neighborhoods and effectively decrease attribute fluctuation’s negative influence. For evaluating our method, we compare it with four traditional methods (MIM, MIFS, CIFE, and mRMR) and two state-of-art methods (DCSF and MRMD) on 17 datasets. Results show that our method has the highest average accuracies on four classifiers and the most times of the highest accuracy. In most cases, using ANP can promote mutual information method’s performance. Additionally, we also compare the accuracy of the selected features set to the accuracy of the full feature set, our method owns higher average accuracy than the full feature set, which indicates that feature selection indeed promotes the prediction accuracy.

In the future, we need to explore the more accurate method to select the key features and the more effective partition method to eliminate attribute fluctuation.
Fig. 9. ANP vs. FP
Acknowledgements

This work is supported in part by the scholarship from China Scholarship Council (CSC) under the Grant CSC N°201806280502. We also thank the experts who have contributed to the datasets.

References

Ayesha, S., Hanif, M. K., & Talib, R. (2020). Overview and comparative study of dimensionality reduction techniques for high dimensional data. Information Fusion, 59, 44–58.

Battiti, R. (1994). Using mutual informations for selecting features in supervised neural net learning. IEEE Transactions on neural networks, 5, 537–550.

Dua, D., & Graff, C. (2019). Uci machine learning repository irvine, ca: University of california, school of information and computer science, 2017.

Fleuret, F. (2004). Fast binary feature selection with conditional mutual information. Journal of Machine learning research, 5.

Gao, W., Hu, L., & Zhang, P. (2018). Class-specific mutual information variation for feature selection. Pattern Recognition, 79, 328–339.

Gao, W., Hu, L., & Zhang, P. (2020). Feature redundancy term variation for mutual information-based feature selection. Applied Intelligence, 50, 1272–1288.

Garcia, S., Luengo, J., Sáez, J. A., Lopez, V., & Herrera, F. (2012). A survey of discretization techniques: Taxonomy and empirical analysis in supervised learning. IEEE Transactions on Knowledge and Data Engineering, 25, 734–750.

Hoque, N., Bhattacharyya, D. K., & Kalita, J. K. (2014). Mifs-nd: A mutual information-based feature selection method. Expert Systems with Applications, 41, 6371–6385.

Hu, Q., Yu, D., Liu, J., & Wu, C. (2008). Neighborhood rough set based heterogeneous feature subset selection. Information sciences, 178, 3577–3594.

Hu, Q., Zhang, L., Zhang, D., Pan, W., An, S., & Pedrycz, W. (2011). Measuring relevance between discrete and continuous features based on neighborhood mutual information. Expert Systems with Applications, 38, 10737–10750.

Hu, Q.-H., & Yu, D.-R. (2009). Neighborhood entropy. In 2009 International Conference on Machine Learning and Cybernetics (pp. 1776–1782). IEEE volume 3.

Kamalov, F., & Thabtah, F. (2017). A feature selection method based on ranked vector scores of features for classification. Annals of Data Science, 4, 483–502.

Kwak, N., & Choi, C.-H. (2002). Input feature selection for classification problems. IEEE transactions on neural networks, 13, 143–159.

Lewis, D. D. (1992). Feature selection and feature extraction for text categorization. In Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992.

Lin, D., & Tang, X. (2006). Conditional infomax learning: An integrated framework for feature extraction and fusion. In European conference on computer vision (pp. 68–82). Springer.

Mariello, A., & Battiti, R. (2018). Feature selection based on the neighborhood entropy. IEEE transactions on neural networks and learning systems, 29, 6313–6322.
Pawlak, Z. (1982). Rough sets. *International journal of computer & information sciences, 11*, 341–356.

Pawlak, Z., Grzymala-Busse, J., Slowinski, R., & Ziarko, W. (1995). Rough sets. *Communications of the ACM, 38*, 88–95.

Peker, N., & Kubat, C. (2021). Application of chi-square discretization algorithms to ensemble classification methods. *Expert Systems with Applications, 185*, 115540.

Peng, H., Long, F., & Ding, C. (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*, 1226–1238.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal, 27*, 379–423.

Sharmin, S., Shoyaib, M., Ali, A. A., Khan, M. A. H., & Chae, O. (2019). Simultaneous feature selection and discretization based on mutual information. *Pattern Recognition, 91*, 162–174.

Sun, L., Zhang, X., Qian, Y., Xu, J., & Zhang, S. (2019). Feature selection using neighborhood entropy-based uncertainty measures for gene expression data classification. *Information Sciences, 502*, 18–41.

Tahan, M. H., & Asadi, S. (2018). Emdid: Evolutionary multi-objective discretization for imbalanced datasets. *Information Sciences, 432*, 442–461.

Thabtah, F., & Abdelhamid, N. (2016). Deriving correlated sets of website features for phishing detection: a computational intelligence approach. *Journal of Information & Knowledge Management, 15*, 1650042.

Thabtah, F., Kamalov, F., Hammoud, S., & Shahamiri, S. R. (2020). Least loss: A simplified filter method for feature selection. *Information Sciences, 534*, 1–15.

Tran, B., Xue, B., & Zhang, M. (2017). A new representation in pso for discretization-based feature selection. *IEEE Transactions on Cybernetics, 48*, 1733–1746.

Tsai, C.-F., & Chen, Y.-C. (2019). The optimal combination of feature selection and data discretization: An empirical study. *Information Sciences, 505*, 282–293.

Wang, G., Lauri, F., & El Hassani, A. H. (2021). A study of dimensionality reduction’s influence on heart disease prediction. In *2021 12th International Conference on Information, Intelligence, Systems & Applications (IISA)* (pp. 1–6). IEEE.

Wulandari, C. P., Ou-Yang, C., & Wang, H.-C. (2019). Applying mutual information for discretization to support the discovery of rare-unusual association rule in cerebrovascular examination dataset. *Expert Systems with Applications, 118*, 52–64.

Xiao, L., Wang, C., Dong, Y., & Wang, J. (2019). A novel sub-models selection algorithm based on max-relevance and min-redundancy neighborhood mutual information. *Information Sciences, 486*, 310–339.

Yang, H. H., & Moody, J. E. (1999). Data visualization and feature selection: new algorithms for nongaussian data. In *NIPS*. Citeseer volume 12.

Zhang, R., Nie, F., Li, X., & Wei, X. (2019). Feature selection with multi-view data: A survey. *Information Fusion, 50*, 158–167.