Mathematical Biosciences and Engineering

http://www.aimspress.com/journal/MBE

MBE, 18(1): 305–327.
DOI: 10.3934/mbe.2021016
Received: 06 July 2020
Accepted: 09 November 2020
Published: 30 November 2020

Research article

Feature selection based on fuzzy joint mutual information maximization

Omar A. M. Salem [1,2], Feng Liu [1,*], Ahmed Sobhy Sherif [2], Wen Zhang [3] and Xi Chen [1,*]

[1] School of Computer Science, Wuhan University, Wuhan 430072, China
[2] Faculty of Computers and Informatics, Suez Canal University, Ismailia 41522, Egypt
[3] College of informatics, Huazhong Agricultural University, Wuhan 430070, China

* **Correspondence:** Email: fliuwhu@whu.edu.cn, robertcx@whu.edu.cn.

Abstract: Nowadays, real-world applications handle a huge amount of data, especially with high-dimension features space. These datasets are a significant challenge for classification systems. Unfortunately, most of the features present are irrelevant or redundant, thus making these systems inefficient and inaccurate. For this reason, many feature selection (FS) methods based on information theory have been introduced to improve the classification performance. However, the current methods have some limitations such as dealing with continuous features, estimating the redundancy relations, and considering the outer-class information. To overcome these limitations, this paper presents a new FS method, called Fuzzy Joint Mutual Information Maximization (FJMIM). The effectiveness of our proposed method is verified by conducting an experimental comparison with nine conventional and state-of-the-art feature selection methods. Based on 13 benchmark datasets, experimental results confirm that our proposed method leads to promising improvement in classification performance and feature selection stability.

Keywords: mutual information; fuzzy sets; fuzzy mutual information; feature selection; classification systems

1. Introduction

Recently, classification systems have a wide range in many fields such as text classification, intrusion detection, bio-informatics, and image retrieval [1]. Unfortunately, a huge amount of data which may include irrelevant or redundant features is one of the main challenges of these systems. The negative effect of these undesirable features reduces the classification performance [2]. For this reason, reducing the number of features by finding an effective subset of features is an important task in classification systems [2]. Feature reduction has two techniques: feature selection and feature extraction [3]. Both reduce a high-dimensional dataset into a representative feature subset of

low-dimensional. Feature extraction is effective when the original features fail to discriminate the classes [4], but it requires extra computation. Moreover, it changes the true meaning of the original features. In contrast, the feature selection preserves the true meaning of the selected features, which is important for some classification systems [5]. Furthermore, the result of FS is more understandable for domain experts [6].

FS tries to find the best feature subset which represents the dataset well and improves the performance of classification systems [7]. It can be classified into three approaches [6]: wrapper, embedded, and filter. According to an evaluation strategy, wrapper and embedded are called classifier-dependent approaches, while filter is called classifier-independent approach [8]. In this paper, we use the filter approach according to its advantages over wrapper or embedded approaches in terms of efficiency, simplicity, scalability, practicality, and classifier-independently [6, 9]. Filter approach is a pre-processing task which finds the highly ranked features to be the input of classification systems [7, 10]. There are two criteria to rank features: feature relevance, and feature redundancy [11]. Feature relevance is related to how features discriminate different classes, while feature redundancy is related to how features share the same information of each other [12]. To define these criteria, filter approach uses many weighting functions which rank features based on their significance [10] such as correlation [13], mutual information (MI) [14]. MI overcomes the weakness of correlation, whereas, correlation is suitable only for linear relationship and numerical features [1]. MI is suitable for any kind of relationship such as linear and non-linear. Moreover, MI deals with both numerical and categorical features [1].

Although MI has been widely used in many methods to find the best feature subset that maximizes the relevancy between the candidate feature and class label, and minimizes the redundancy between the candidate feature and pre-selected features [15]. The main limitations of these methods are: (1) difficult to indicate the best candidate features with the same new classification information [16], (2) difficult to deal with continuous features without information loss [17], and (3) consider the inner-class information only [18]. In this paper, we integrate fuzzy concept with mutual information to propose a new FS method called Fuzzy Joint Mutual Information Maximization (FJMIM). The fuzzy concept helps the proposed method to exploit all possible information of data where it can deal with any numerical data and extract the inner and outer-class information. Moreover, the objective function of FJMIM can overcome the feature overestimation problem which happens when the candidate feature be completely correlated with some of pre-selected features and does not depend on the majority of the subset at the same time [8].

The rest of this paper is organized as follows: Section 2 presents the basic measures of fuzzy information theory. Then, we present the proposed method in section 3. After that, the experiment design was presented in section 4, followed by the results and discussion in section 5. Finally, section 6 concludes the paper.

2. Basic measures of fuzzy information theory

For the purpose of measuring the significance of features, information theory introduced many information measures such as entropy, and mutual information. To enhance these measures, fuzzy concept is used to estimate new extensions of information measures based on fuzzy equivalence relations such as fuzzy entropy, and fuzzy mutual information [19, 20]. Fuzzy entropy measures the

average amount of uncertainty of fuzzy relation in order to estimate its discriminative power, while fuzzy mutual information measures the shared amount of information between two fuzzy relations. In the following, we present the basic measures of fuzzy information theory:

Given a dataset $D = F \cup C$, where F is a set of n features, and C is the class label. Let $\bar{F} = \{a_1, a_2, \ldots, a_m\}$ be a feature of m samples, where $\bar{F} \in F$. Let S is the feature subset with d of selected features, and the remaining set is $\{F - S\}$, where $\bar{F}_f \in F - S$ and $\bar{F}_s \in S$. Based on the fuzzy equivalence relation $R_{\bar{F}}$ on $\bar{F}$, the feature $\bar{F}$ can be represented by the relation matrix $M(R_{\bar{F}})$.

$$M(R_{\bar{F}}) = \begin{pmatrix} r_{11} & r_{12} & \ldots & r_{1m} \\ r_{21} & \ldots & \ldots & r_{2m} \\ \ldots & \ldots & \ldots & \ldots \\ r_{m1} & r_{m2} & \ldots & r_{mm} \end{pmatrix} \tag{2.1}$$

where $r_{ij} = R_{\bar{F}}(a_i, a_j)$ is the fuzzy equivalence relation between two samples a_i and a_j.

In this paper, the used fuzzy equivalence relation between two elements a_i and a_j is defined as [21]:

$$R_{\bar{F}}(a_i, a_j) = \exp - \|a_i - a_j\| \tag{2.2}$$

Fuzzy equivalence class of sample a_i on $R_{\bar{F}}$ can be defined as:

$$[a_i]_{R_{\bar{F}}} = \frac{[r_{i1}]}{a_1} + \frac{[r_{i2}]}{a_2} + \cdots + \frac{[r_{im}]}{a_m} \tag{2.3}$$

Fuzzy entropy of feature $\bar{F}_1$ based on fuzzy equivalence relation is defined as:

$$H(\bar{F}) = \frac{1}{m} \sum_{i=1}^{m} \log \frac{m}{|[a_i]_{R_{\bar{F}}}|} \tag{2.4}$$

where $|[a_i]_{R_{\bar{F}}}| = \sum_{i=1}^{m} r_{ij}$.

Let $\bar{F}_1$ and $\bar{F}_2$ be two features of F, fuzzy joint entropy of $\bar{F}_1$ and $\bar{F}_2$ is defined as:

$$\begin{aligned} H(\bar{F}_1, \bar{F}_2) &= H(R_{\bar{F}_1}, R_{\bar{F}_2}) \\ &= \frac{1}{m} \sum_{i=1}^{m} \log \frac{m}{|[a_i]_{R_{\bar{F}_1}} \cap [a_i]_{R_{\bar{F}_2}}|} \end{aligned} \tag{2.5}$$

Fuzzy conditional entropy of $\bar{F}_1$ given $\bar{F}_2$ is defined as

$$\begin{aligned} H(\bar{F}_1 | \bar{F}_2) &= H(R_{\bar{F}_1} | R_{\bar{F}_2}) \\ &= \frac{1}{m} \sum_{i=1}^{m} \log \frac{|[a_i]_{R_{\bar{F}_2}}|}{|[a_i]_{R_{\bar{F}_1}} \cap [a_i]_{R_{\bar{F}_2}}|} \end{aligned} \tag{2.6}$$

Fuzzy Mutual information between two features $\bar{F}_1$ and $\bar{F}_2$ is defined as:

$$\begin{aligned} I(\bar{F}_1; \bar{F}_2) &= I(R_{\bar{F}_1}; R_{\bar{F}_2}) \\ &= \frac{1}{m} \sum_{i=1}^{m} \log \frac{m|[a_i]_{R_{\bar{F}_1}} \cap [a_i]_{R_{\bar{F}_2}}|}{|[a_i]_{R_{\bar{F}_1}} . [a_i]_{R_{\bar{F}_2}}|} \end{aligned} \tag{2.7}$$

Fuzzy conditional mutual information between feature $\bar{F}_1$ and $\bar{F}_2$ given class C is defined as:

$$I(\bar{F}_1;\bar{F}_2|C) = H(\bar{F}_1|C) + H(\bar{F}_2|C) - H(\bar{F}_1,\bar{F}_2|C) \tag{2.8}$$

Fuzzy joint mutual information between two features $\bar{F}_1$, $\bar{F}_2$ and class C is defined as:

$$I(\bar{F}_1,\bar{F}_2;C) = I(\bar{F}_1;C) + I(\bar{F}_2;C|\bar{F}_1) \tag{2.9}$$

Fuzzy interaction information between among $\bar{F}_1$, $\bar{F}_2$ and C is defined as:

$$I(\bar{F}_1;\bar{F}_2;C) = I(\bar{F}_1;C) + I(\bar{F}_2;C) - I(\bar{F}_1,\bar{F}_2;C) \tag{2.10}$$

3. Proposed feature selection method

In this section, we presented the general theoretical frameworks of different feature selection methods based on mutual information. Then, we studied the limitation of previous work. Finally, we introduced the proposed method.

3.1. Feature selection based on mutual information

Brown et al. [22] studied the exist feature selection methods based on MI and analyzed the different criteria to propose the following theoretical framework of these methods.

$$J(\bar{F}_f) = I(\bar{F}_f;C) - \beta \sum_{\bar{F}_s \in S} I(\bar{F}_f;\bar{F}_s) + \gamma \sum_{\bar{F}_s \in S} I(\bar{F}_f;\bar{F}_s|C) \tag{3.1}$$

This framework is a linear combination of three terms: relevance, redundancy, and conditional that measures the individual predictive power of the feature, the unconditional relation, and the class-conditional relation, respectively. The criteria of different feature selection based on MI depends on the value of β and γ. MIM ($\beta = \gamma = 0$) [23] is the simplest FS method based on MI. It considers only the relevance relation only. However, It may suffer from the redundant features. MIFS ($\gamma = 0$) [24] introduced two criteria to estimate the feature relevance and redundancy. An extension of MIFS, called MIFS-U [24] is proposed to improve the redundancy term of MIFS by considering the uniform distribution of the information. However, Both MIFS and MIFS-U still require an input parameter β. To avoid this limitation, MRMR ($\beta = \frac{1}{|S|}$, $\gamma = 0$) [25] introduced the mean of the redundancy term as automatic value to the input parameter (β). JMI ($\beta = \gamma = \frac{1}{|S|}$) [26] extended MRMR to extract the benefit of conditional term. In addition, Brown et al. [22] introduced also a similar non-linear framework to represent some methods as CMIM method [27]. According to [22], CMIM can be written as:

$$\begin{aligned} J_{cmim} &= \min_{\bar{F}_s \in S}[I(\bar{F}_f;C|\bar{F}_s)] \\ &= I(\bar{F}_f;C) - \max_{\bar{F}_s \in S}[I(\bar{F}_f;\bar{F}_s) - I(\bar{F}_f;\bar{F}_s|C)] \end{aligned} \tag{3.2}$$

The reason of the non-linear relation on CMIM returns to the using of *max* operation. Similar to CMIM, JMIM [8] introduces a non-linear relation as follows:

$$\begin{aligned} J_{jmim} &= \min_{\bar{F}_s \in S}[I(\bar{F}_s;C) + I(\bar{F}_f;C|\bar{F}_s)] \\ &= I(\bar{F}_f;C) - \max_{\bar{F}_s \in S}[I(\bar{F}_f;\bar{F}_s) - I(\bar{F}_f;\bar{F}_s|C) - I(\bar{F}_s;C)] \end{aligned} \tag{3.3}$$

3.2. Limitation of previous work

Although MI has been widely used in many feature selection methods such as MIFS [24], JMI [26], mRMR [25], DISR [28], IGFS [29], NMIFS [30] and MIFS-ND [31]. These methods suffer from the overestimation of the feature significance problem [8]. For this reason, Bennasar et al. [8] proposed JMIM method to address the overestimation of the feature significance problem. However, it may fail to select the best candidate features if they have the same new classification information. To illustrate this problem, Figure 1 shows the FS scenario, where $\bar{F}_1$ and $\bar{F}_2$ are two candidate features, $\bar{F}_s$ is the pre-selected feature subset, and C is the class label. $\bar{F}_1$ is partially redundant with $\bar{F}_s$, while $\bar{F}_2$ is independent to $\bar{F}_s$. Suppose that $\bar{F}_1$ and $\bar{F}_2$ have the same new classification information $I(\bar{F}_1;C|\bar{F}_s)$= (area 3) and $I(\bar{F}_2;C|\bar{F}_s)$= (area 5) respectively. In this case, JMIM may fail to indicate the best feature where $I(\bar{F}_1,\bar{F}_s;C)$ and $I(\bar{F}_1,\bar{F}_s;C)$ are equal.

Unfortunately, JMIM also shares some limitations with the previous methods. Firstly, it can not directly estimate MI between continuous features [17]. To address this limitation, there are two methods were introduced. One is to estimate MI based on Parzen window [18], but it is inefficient in high-dimensional feature spaces with spare samples [17]. Moreover, its performance depends on the used window function which requires a window width parameter [32]. The other one is to discretize continuous features before estimating MI [33], but it may cause information loss [34]. Secondly, JMIM depends only on inner-class information without considering outer-class information [18].

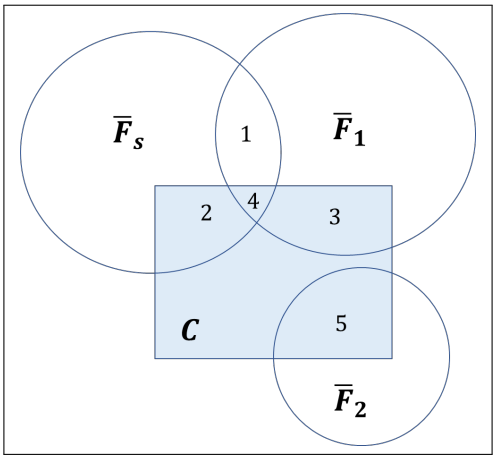


Figure 1. Venn diagram presents the feature selection scenario where the two candidate features $\bar{F}_1$ and $\bar{F}_2$ have the same new classification information.

3.3. Fuzzy Joint Mutual Information (FJMIM)

Motivated by the previous limitation of JMIM, we proposed a new FS method, called Fuzzy Joint Mutual Information Maximization (FJMIM). Both of FJMIM and JMIM depends on "maximum of the minimum" approach. The main difference is that JMIM maximizes the joint mutual information of the candidate feature and pre-selected feature subset with class, whereas FJMIM maximizes the joint mutual information of the candidate feature and pre-selected feature subset with class without considering the class-relevant redundancy. To illustrate the difference in Figure 1, JMIM depends on

the union of areas 2, 4, and 3, while FJMIM depends on the union of areas 2 and 3. The proposed method discarded the class-relevant redundancy (area 4) because it can reduce the predictive ability of the feature subset when a candidate feature is selected [15]. On the other hand, integrating fuzzy concept with MI has many benefits. Firstly, using a fuzzy concept helps to deal directly with continuous features. Furthermore, it enables MI to take the advantages of inner and outer-class information [35]. Moreover, FS methods based on fuzzy concept are more robust toward any change of the data than methods based on probability concept [36].

According to FJMIM, the candidate feature $\bar{F}_f$ must satisfy the following condition

$$\bar{F}_f = arg \max_{\bar{F}_f \in F-S} (\min_{\bar{F}_s \in S} (I(\bar{F}_f, \bar{F}_s; C) - I(\bar{F}_f; \bar{F}_s; C))) \tag{3.4}$$

FJMIM also can be written according to the non-linear framework as follows:

$$\bar{F}_f = 2 * I(\bar{F}_f; C) - \max_{\bar{F}_s \in S} [I(\bar{F}_f; \bar{F}_s) - I(C; \bar{F}_s) - I(\bar{F}_f; \bar{F}_s | C) - I(\bar{F}_f; C | \bar{F}_s)] \tag{3.5}$$

The proposed method can be summarized to find the best feature subset of size d as follows:

Input: F is a set of n features, C is the class label, and d is the number of selected features.
Step 1: Initialize the empty selected feature subset S.
Step 2: Update the selected feature set S and the feature set F.

- **Step 2.1:** Compute $I(\bar{F}; C)$ for all $\bar{F}$ in the feature set F.
- **Step 2.2:** Add the feature $\bar{F}$ that maximizes $I(\bar{F}; C)$ to the selected feature set S.
- **Step 2.3:** Remove the feature $\bar{F}$ from the feature set F.

Step 3: Repeat until $|S| = d$

- **Step 3.1:** Add the feature $\bar{F}$ that satisfies
  $arg \max_{\bar{F}_f \in F-S} (\min_{\bar{F}_s \in S} (I(\bar{F}_f, \bar{F}_s; C) - I(\bar{F}_f; \bar{F}_s; C)))$
  to the selected feature set S.
- **Step 3.2:** Remove the feature $\bar{F}$ from the feature set F.

Output: Return the selected feature set S.

4. Experiment

Success of FS methods depends on different criteria such as classification performance, and stability [11]. Consequently, we design the experiment based on these criteria (Figure 2). To clarify our improvement, we compared our proposed method FJMIM, with four conventional methods (CMIM [27], JMI [26], QPFS [37], Relief [38]) and five state-of-the-art methods (CMIM3 [39], JMI3 [39], JMIM [8], MIGM [40], and WRFS [41]). The compared methods can be divided into two groups: FS based on fuzzy concept and FS based on probability concept. For the methods which depend on probability concept, data discretization is required as a pre-processing step prior to the FS process. So, the continuous features are transformed into ten bins using EqualWidth discretization [42]. Then, we selected feature subset from all methods based on threshold which is defined as the median position of the ranked features (or the nearest integer position when the number of ranked features is even).

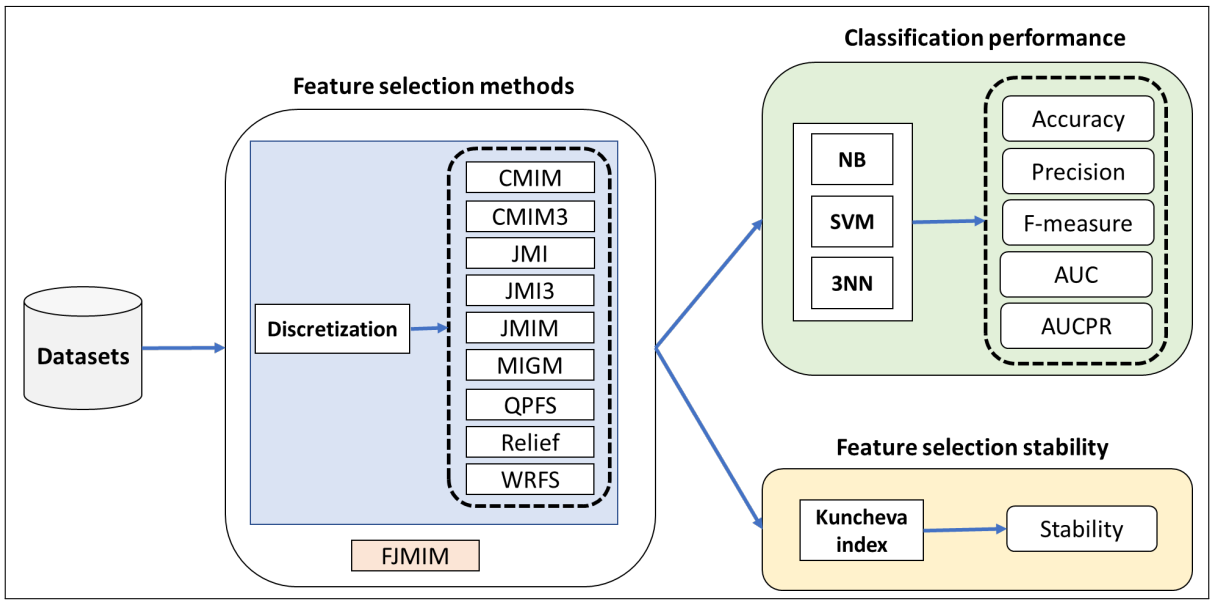


Figure 2. Experiment design of the proposed method Fuzzy Joint Mutual Information (FJMIM): firstly, a discretization pre-processing is applied before FS methods based on probability concept. Then, the FS methods are evaluated according to classification performance, and feature stability.

4.1. Evaluation criteria

4.1.1. Classification performance

FS plays an important role to improve the classification performance. There are many measures to evaluate classification performance such as classification accuracy, precision, F-measure, area-under-ROC (AUC), and area-under-PRC (AUCPR) [43]. To clarify our improvement, popular classifiers were used in this study such as Naive Bayes (NB), Support Vector Machine (SVM), and 3-Nearest Neighbors (KNN) [44]. The average classification performance measures were computed by 10-fold cross-validation approach [45].

4.1.2. Stability

Another important evaluation criterion for FS is stability. FS stability measures the impact of any change in the input data on FS result [46]. In this study, we measure the impact of noise on the selected feature subset. Firstly, we produce the noise using standard deviation and the normal distribution of each feature [47]. Then, we injected 10% of the data by adding noise. After that, we repeated this step ten times. Each time produces a different sequence of selected features. Finally, we computed the stability of each method using Kuncheva stability index [48].

4.2. Datasets

Our experiment was conducted using 13 datasets from UCI machine learning repository [49]. Table 1 presents a brief description about these datasets.

Table 1. Datasets Description.

No.	Datasets	Instances	Features	Classes
1	Acute Inflammations	120	6	2
2	Arrhythmia	452	279	13
3	Blogger	100	6	2
4	Diabetic Retinopathy Debrecen (DRD)	1151	20	2
5	Hayes-Roth	160	5	3
6	Indian Liver Patient Dataset (ILPD)	583	10	2
7	Lenses	24	4	3
8	Lymphography	148	18	4
9	Congressional Voting Records (CVR)	435	16	2
10	Sonar	208	60	2
11	Thoracic Surgery	470	17	2
12	Wilt	4889	6	2
13	Zoo	101	17	7

5. Results and discussion

5.1. Classification performance

1) Accuracy: A paired two-tailed t-test is employed between FJMIM and other compared methods. The notations ([=], [+], and [-]) indicate the statistically significant (5%) that the proposed method (equals, wins, and losses) other methods. According to NB classifier, FJMIM achieved the maximum average accuracy with score 78.02%, while Relief achieved the minimum average accuracy with score 75.34% (Table 2). The proposed method outperformed compared methods in the range from 0.06 to 2.68%. In SVM classifier, FJMIM outperformed other methods with score 80.03%, while Relief achieved the minimum average accuracy with score 77.33% (Table 3). The proposed method outperformed compared methods in the range from 0.69 to 2.7%. Similarly with KNN classifier, FJMIM kept the maximum average accuracy by 81.45%, while Relief achieved the minimum average accuracy with score 77.98% (Table 4). The proposed method outperformed compared methods in the range from 0.57 to 3.47%. Across all datasets, Figure 3(a) shows the distribution of the average accuracy values of all used classifiers. In a detailed box-plot, the box represents upper and lower quartiles, while the black circle represents the median. The box-plot confirms that FJMIM is more consistent and outperformed other compared methods. Figure 3(b) shows the average accuracy of the three used classifiers. FJMIM achieved the best accuracy, followed by QPFS, JMI3, both of JMI and CMIM, both of CMIM3 and JMIM, MIGM, WRFS, and Relief respectively. The proposed method outperformed compared methods in the range from 0.6 to 2.9%.

2) Precision: Figure 4 shows the precision results of NB, SVM, KNN, and their average. FJMIM achieved the highest precision, while Relief achieved the lowest precision. The proposed method

Table 2. Classification accuracy using NB Classifier, FJMIM achieved the highest average accuracy.

Dataset	CMIM	CMIM3	JMI	JMI3	JMIM	MIGM	QPFS	Relief	WRFS	FJMIM
Acute Inflammations	100±0[=]	100±0[=]	100±0[=]	100±0[=]	100±0[=]	99.75±2.5[=]	99.75±2.5[=]	100±0[=]	90.83±8.5[+]	100±0
Arrhythmia	60.64±7.01[=]	59.22±6.17[=]	59.42±6.49[=]	59.58±6.58[=]	60.62±6.93[=]	58.98±6.39[=]	63.7±6.57[-]	64.94±6.33[-]	59.31±6.3[=]	59.74±6.43
Blogger	61.9±10.7[=]	61.9±10.7[=]	61.9±10.7[=]	64.9±10.2[=]	61.9±10.7[=]	61.9±10.7[=]	61.9±10.7[=]	65.4±7.84[=]	66.1±9.09[=]	65.9±10.26
DRD	57.59±4.95[=]	60.66±4.69[-]	60.3±5.04[-]	60.6±4.58[-]	57.59±4.95[=]	60.3±5.04[-]	61.19±4.49[-]	60.61±5.06[-]	57.47±4.98[=]	57.56±4.83
Hayes-Roth	45.81±10.77[=]	45.81±10.77[=]	45.81±10.77[=]	45.81±10.77[=]	45.81±10.77[=]	45.81±10.77[=]	49.37±10.53[=]	46.31±11.48[=]	45.81±10.77[=]	49.37±10.53
ILPD	68.99±5.68[=]	67.33±5.87[=]	68.99±5.68[=]	67.33±5.87[=]	68.99±5.68[=]	68.99±5.68[=]	68.73±5.64[=]	64.96±6.45[+]	70.11±6.74[=]	69.54±5.75
Lenses	86.83±21.88[=]	86.83±21.88[=]	86.83±21.88[=]	86.83±21.88[=]	86.83±21.88[=]	86.83±21.88[=]	86.83±21.88[=]	54.83±30.73[+]	86.83±21.88[=]	86.83±21.88
Lymphography	79.02±10.57[=]	82.05±9.59[=]	79.02±10.57[=]	77.93±10.87[=]	78.27±10.98[=]	77.25±11.38[=]	79.02±10.57[=]	82.61±9.77[=]	76.91±11.64[+]	81.5±10.29
CVR	94.3±3.15[=]	94.32±3.12[=]	93.75±3.57[=]	93.75±3.57[=]	93.75±3.57[=]	93.75±3.57[=]	94.32±3.26[=]	93.82±3.28[=]	93.52±3.18[=]	93.75±3.57
Sonar	75.05±8.33[=]	75.42±9.08[=]	75.75±8.85[=]	76.23±9.99[=]	72.2±9.93[=]	72.14±9.54[+]	76.42±8.69[=]	74.25±8.33[=]	77.91±8.74[=]	77±8.98
Thoracic Surgery	83.83±3.1[=]	83.38±3.09[=]	83.38±3.09[=]	83.85±2.24[=]	83.38±3.09[=]	84.64±1.83[=]	83.55±3.08[=]	83.06±2.77[=]	83.85±2.74[=]	83.38±3.09
Wilt	94.61±0.06[=]	94.61±0.06[=]	94.61±0.06[=]	94.61±0.06[=]	94.61±0.06[=]	94.61±0.06[=]	94.61±0.06[=]	94.61±0.06[=]	94.61±0.06[=]	94.61±0.06
Zoo	95.15±5.74[=]	96.05±5.6[=]	96.03±5.66[=]	96.03±5.66[=]	96.03±5.66[=]	96.93±5.42[=]	94.08±6.6[=]	94.08±6.6[=]	92.96±7.05[=]	95.05±5.87
Average	77.21	77.51	77.37	77.50	76.92	77.07	77.96	75.34	76.63	78.02

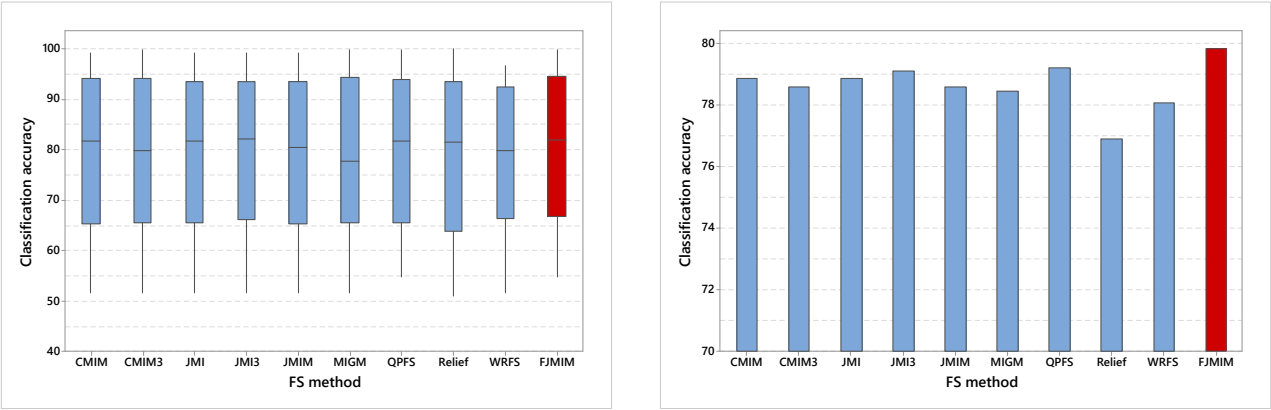
Table 3. Classification accuracy using SVM Classifier, FJMIM achieved the highest average accuracy.

Dataset	CMIM	CMIM3	JMI	JMI3	JMIM	MIGM	QPFS	Relief	WRFS	FJMIM
Acute Inflammations	98.33±3.35[=]	99.67±1.64[=]	98.33±3.35[=]	98.33±3.35[=]	98.33±3.35[=]	99.42±2.44[=]	99.42±2.44[=]	100±0[+]	89.92±9.5[+]	99.42±2.44
Arrhythmia	66.09±6.81[=]	64.08±6.11[=]	65.01±5.96[=]	65.7±6.23[=]	66.14±6.61[=]	65.23±6.39[=]	66.82±6.2[=]	67.22±6.05[=]	65.23±6.15[=]	67.11±6.7
Blogger	68±4.02[=]	67.7±4.89[=]	68±4.02[=]	67.4±5.62[=]	68±4.02[=]	68±4.02[=]	67.7±4.89[=]	67.7±4.89[=]	68±4.02[=]	67.7±4.89
DRD	68.03±3.92[=]	67.55±4.07[=]	67.15±3.85[=]	67.03±3.95[=]	68.03±3.92[=]	67.19±3.8[=]	65.1±4.33[+]	61.91±4.51[+]	67.7±4[=]	67.96±4.23
Hayes-Roth	49.94±12.83[=]	49.94±12.83[=]	49.94±12.83[=]	49.94±12.83[=]	49.94±12.83[=]	49.94±12.83[=]	52±11.61[=]	48.75±13.21[=]	49.94±12.83[=]	52±11.61
ILPD	71.36±0.73[=]	71.36±0.73[=]	71.36±0.73[=]	71.36±0.73[=]	71.36±0.73[=]	71.36±0.73[=]	71.36±0.73[=]	71.36±0.73[=]	71.36±0.73[=]	71.36±0.73
Lenses	76.17±22.13[=]	76.17±22.13[=]	76.17±22.13[=]	76.17±22.13[=]	76.17±22.13[=]	76.17±22.13[=]	76.17±22.13[=]	54.5±30.42[+]	76.17±22.13[=]	76.17±22.13
Lymphography	81.57±10.06[=]	80.56±9.09[=]	81.42±10.15[=]	84.55±8.84[=]	82.02±10.48[=]	79.06±10.44[+]	81.57±10.11[=]	80.3±9[=]	74.25±11.45[+]	85.5±8.3
CVR	94.09±3.33[=]	94.46±3.37[=]	94.3±3.34[=]	94.32±3.32[=]	94.3±3.34[=]	94.3±3.34[=]	94±3.44[=]	94.43±3.41[=]	94.37±3.41[=]	94.3±3.34
Sonar	79.34±8.37[=]	80.48±8.05[=]	80.18±8.33[=]	78.46±8.44[=]	80.54±8.43[=]	76.93±8.71[=]	79.91±7.73[=]	82.1±8.07[=]	77.41±8.41[=]	80.67±7.88
Thoracic Surgery	85.11±0[=]	85.11±0[=]	85.11±0[=]	85.11±0[=]	85.11±0[=]	85.11±0[=]	85.11±0[=]	85.11±0[=]	85.11±0[=]	85.11±0
Wilt	97.62±0.57[+]	97.97±0.49[=]	97.97±0.49[=]	97.97±0.49[=]	97.62±0.57[+]	97.97±0.49[=]	94.61±0.06[+]	97.97±0.49[=]	97.62±0.57[+]	97.98±0.49
Zoo	95.75±5.12[=]	93.88±6.72[=]	89.39±5.72[+]	89.49±5.64[+]	89.49±5.64[+]	94.06±6.9[=]	95.27±5.85[=]	93.89±6.96[=]	90.49±7.96[+]	95.06±6.2
Average	79.34	79.15	78.79	78.91	79.00	78.83	79.16	77.33	77.51	80.03

Table 4. Classification accuracy using KNN Classifier, FJMIM achieved the highest average accuracy.

Dataset	CMIM	CMIM3	JMI	JMI3	JMIM	MIGM	QPFS	Relief	WRFS	FJMIM
Acute Inflammations	99.5±1.99[=]	100±0[=]	99.5±1.99[=]	99.5±1.99[=]	99.5±1.99[=]	100±0[=]	100±0[=]	100±0[=]	99.67±1.64[=]	100±0
Arrhythmia	58.27±5.17[+]	57.63±5.17[+]	57.79±5.34[+]	58.32±5.08[+]	58.94±5.13[=]	57.67±5.21[+]	60.73±4.64[=]	65.13±4.43[-]	58.39±5.22[+]	61.11±5.23
Blogger	72.1±11.31[=]	72.1±11.31[=]	72.1±11.31[=]	78.8±10.18[=]	72.1±11.31[=]	72.1±11.31[=]	72.1±11.31[=]	70.3±11.41[=]	76.1±12.05[=]	78.6±11.37
DRD	64.3±4.19[=]	62.62±5.29[=]	63.69±4.58[=]	61.96±4.68[=]	64.3±4.19[=]	63.58±4.56[=]	60.47±4.64[+]	63.51±4.63[=]	64.06±4.4[=]	64.7±4.23
Hayes-Roth	59.19±10.41[=]	59.19±10.41[=]	59.19±10.41[=]	59.19±10.41[=]	59.19±10.41[=]	59.19±10.41[=]	63.12±11.32[=]	57.56±10.82[=]	59.19±10.41[=]	63.12±11.32
ILPD	67.16±5.41[=]	68.22±5.06[=]	67.54±5.44[=]	68.22±5.06[=]	67.16±5.41[=]	67.3±5.42[=]	68.74±5.01[=]	67.84±5.56[=]	66.66±5.25[=]	69.15±5.62
Lenses	86.83±21.88[=]	86.83±21.88[=]	86.83±21.88[=]	86.83±21.88[=]	86.83±21.88[=]	86.83±21.88[=]	86.83±21.88[=]	54.17±30.46[+]	86.83±21.88[=]	86.83±21.88
Lymphography	84.61±7.58[=]	72.61±9.14[+]	84.61±7.58[=]	84.16±9.46[=]	80.99±9.49[=]	76.62±8.87[=]	84.61±7.58[=]	81.44±9.42[=]	79.74±9.36[=]	78.87±8.35
CVR	93.63±3.43[=]	94.09±3.33[=]	95.26±3.01[=]	95.26±3.01[=]	95.26±3.01[=]	95.26±3.01[=]	94.14±3.37[=]	93.67±3.5[=]	94.09±2.95[=]	95.26±3.01
Sonar	85.09±7.43[=]	83.51±7.68[=]	87.74±7.51[=]	85.53±8.83[=]	82.7±9.06[=]	82.4±7.72[+]	86.15±7.24[=]	87.95±6.78[=]	84.17±7.73[=]	87.65±7.27
Thoracic Surgery	80.96±3.45[=]	81.38±3.53[=]	81.38±3.53[=]	83.45±3.05[=]	81.38±3.53[=]	82.64±3.28[=]	83.89±2.95[=]	82.68±3.08[=]	84.85±0.97[-]	81.38±3.53
Wilt	97.69±0.57[+]	98.12±0.55[=]	98.12±0.55[=]	98.12±0.55[=]	97.69±0.57[+]	98.12±0.55[=]	94.44±0.52[+]	98.12±0.55[=]	97.69±0.57[+]	98.12±0.55
Zoo	91.9±7.21[=]	92.06±7.35[=]	92.05±7.24[=]	92.05±7.24[=]	92.05±7.24[=]	91.16±7.48[=]	91.42±6.94[=]	91.42±6.94[=]	90.29±7.45[+]	94.08±6.6
Average	80.09	79.10	80.45	80.88	79.85	79.45	80.51	77.98	80.13	81.45

outperformed other methods in the range from 0.6 to 12.7% based on NB, from 0.3 to 3.3% based on SVM, and from 2 to 8.2% based on KNN. According to the average of all classifiers, FJMIM achieved the highest precision by 71.2%, while Relief kept the lowest precision by 63.1%. The second-best method achieved by CMIM3, followed by JMI3, QPFS, JMI, JMIM, MIGM, CMIM and WRFS. The



(a) Accuracy distribution across all datasets

(b) The average accuracy of all methods

Figure 3. Accuracy result of all compared methods. FJMIM outperformed in all cases.

results of precision distribution are inconsistent as shown in Figure 5. Both FJMIM and QPFS achieved the best distribution. FJMIM achieved the highest upper quartile and median, while QPFS achieved the highest lower quartile and median. In addition, FJMIM shares the highest upper quartile with JMI, JMIM and MIGM.

3) F-measure: Figure 6 shows the F-measure results of the three used classifiers and their average. FJMIM achieved the highest F-measure by 79.8, 71.5 and 84% on NB, SVM and KNN respectively. Relief achieved the lowest F-measure on NB and SVM, while WRFS achieved the lowest score on SVM. The proposed method outperformed other methods in the range from 0.3 to 16.6% based on NB, from 0.1 to 1.4% based on SVM, and from 0.6 to 15.2% based on KNN. According to the average of all classifiers, FJMIM achieved the highest precision by 78.4% and outperformed other methods in the range from 2.5 to 10.8%. The second-best method achieved by JMI3, followed by QPFS, CMIM3, JMI, JMIM, CMIM, MIGM, WRFS and Relief. Figure 7 shows the distribution of F-measure across all datasets. The box-plot confirms the outperformance of FJMIM compared to other methods.

4) AUC: Figure 8 shows the AUC results of the used classifiers and their average. FJMIM achieved the highest AUC on NB and KNN by 83.9 and 85.2%, while it achieved the second-best score on SVM by 74.8%. On the other hand, Relief achieved the lowest AUC on all classifiers. Although MIGM achieved the best AUC on SVM, FJMIM outperformed on the average of all classifiers. The proposed method outperformed other methods in the range from 0.4 to 4.2%. As shown in the box-plot (Figure 9), the proposed method achieved the highest lower quartile and median values. On the other hand, JMI achieved also the highest lower quartile, while CMIM achieved the highest upper quartile.

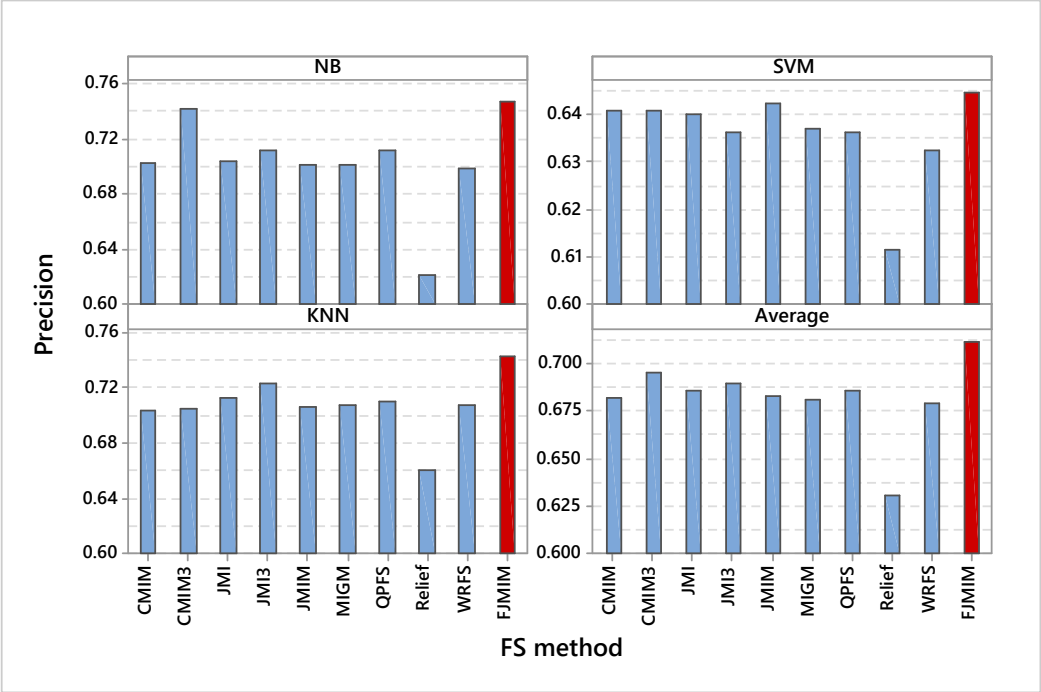


Figure 4. Precision result of NB, SVM, KNN and their average. FJMIM achieved the best result in all cases.

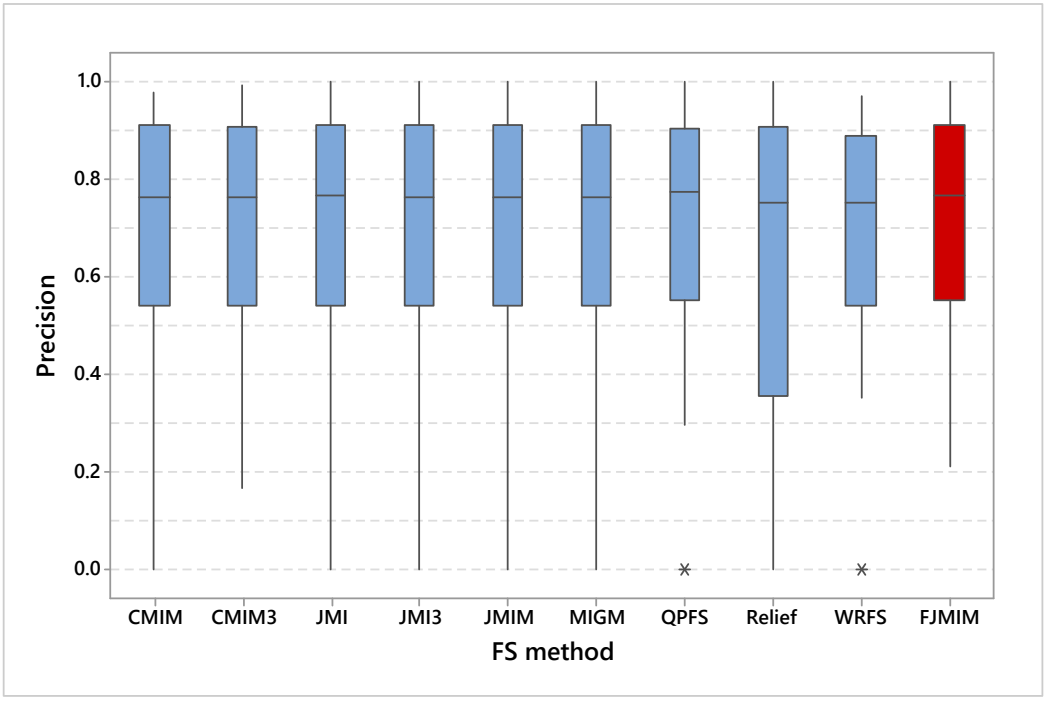


Figure 5. Precision distribution across all datasets.

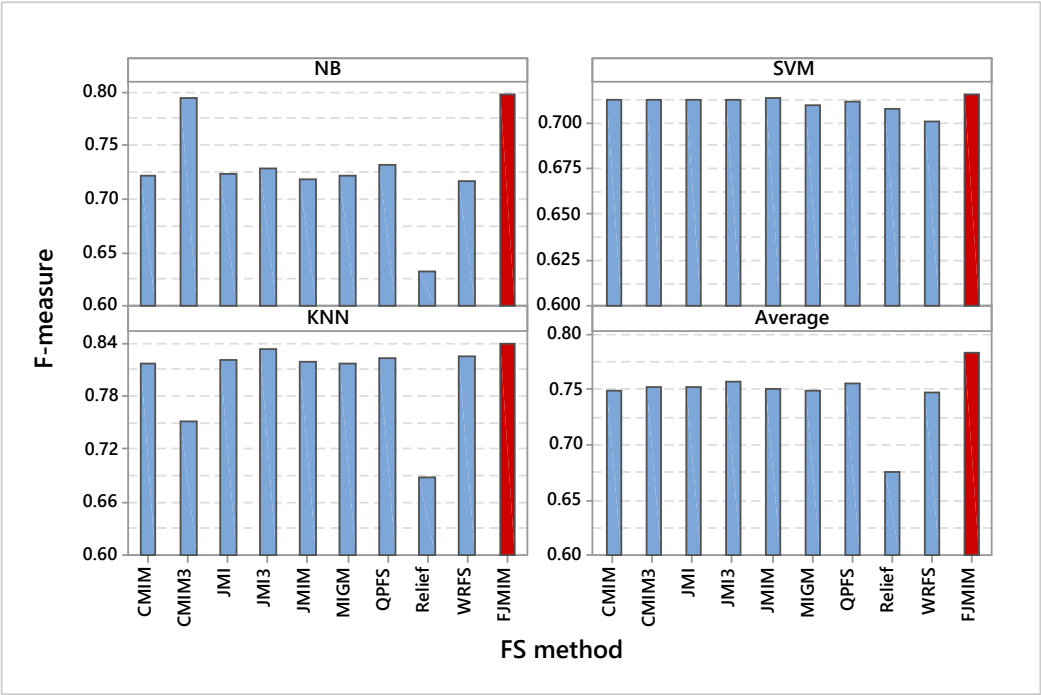


Figure 6. F-measure result of NB, SVM, KNN and their average. FJMIM achieved the best result in all cases.

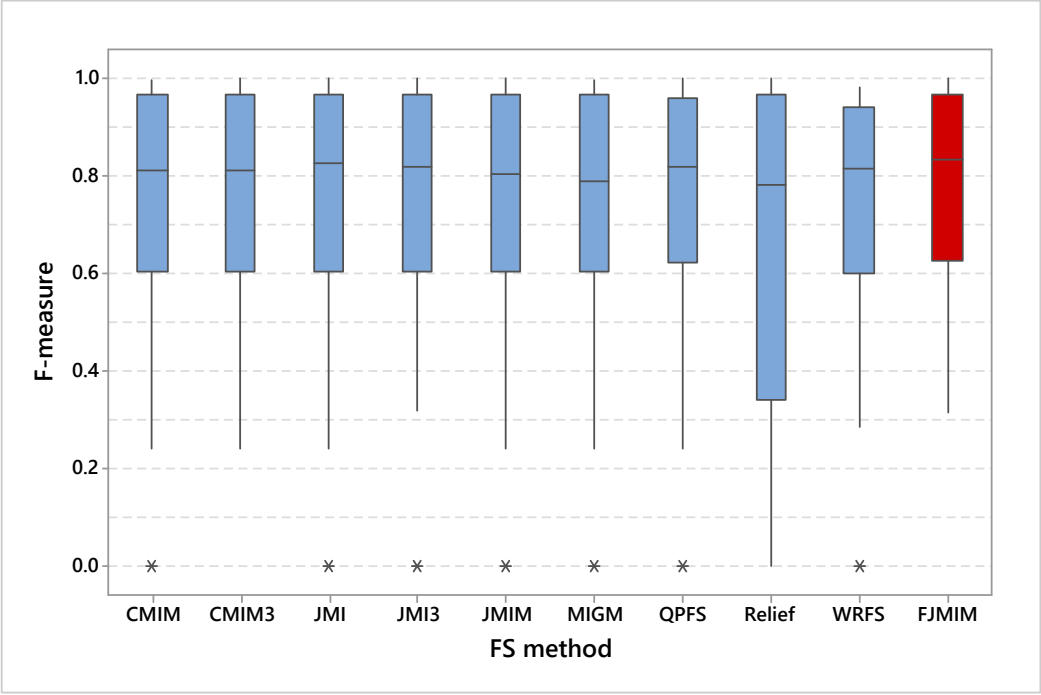


Figure 7. F-measure distribution across all datasets.

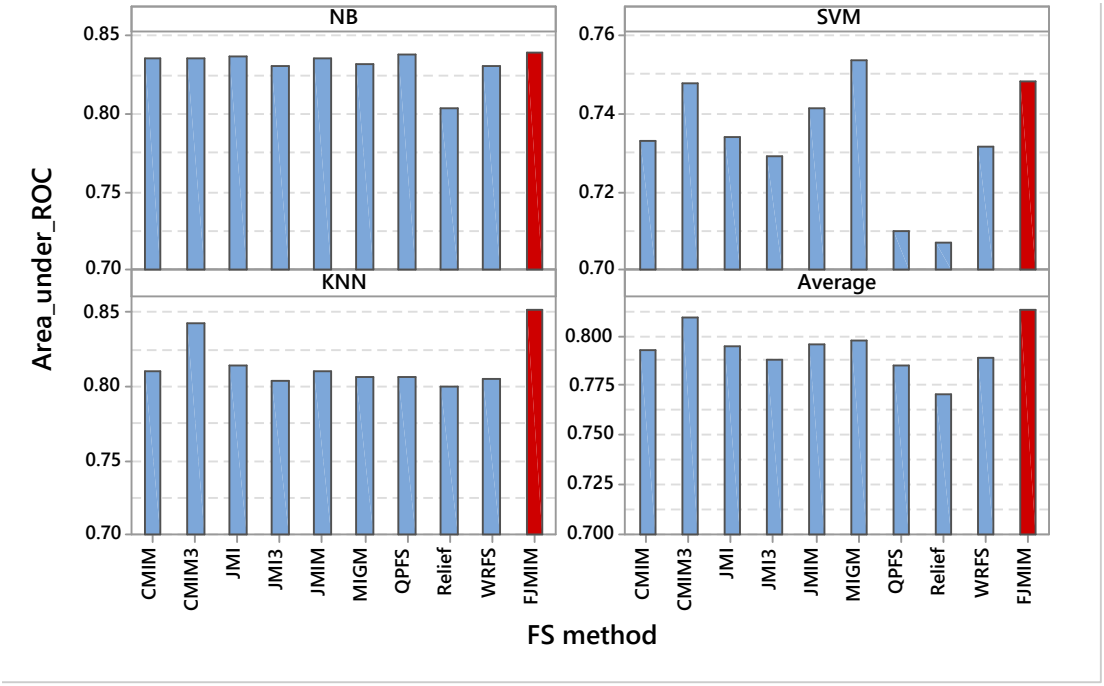


Figure 8. AUC result of NB, SVM, KNN and their average. FJMIM achieved the best result in all cases except SVM.

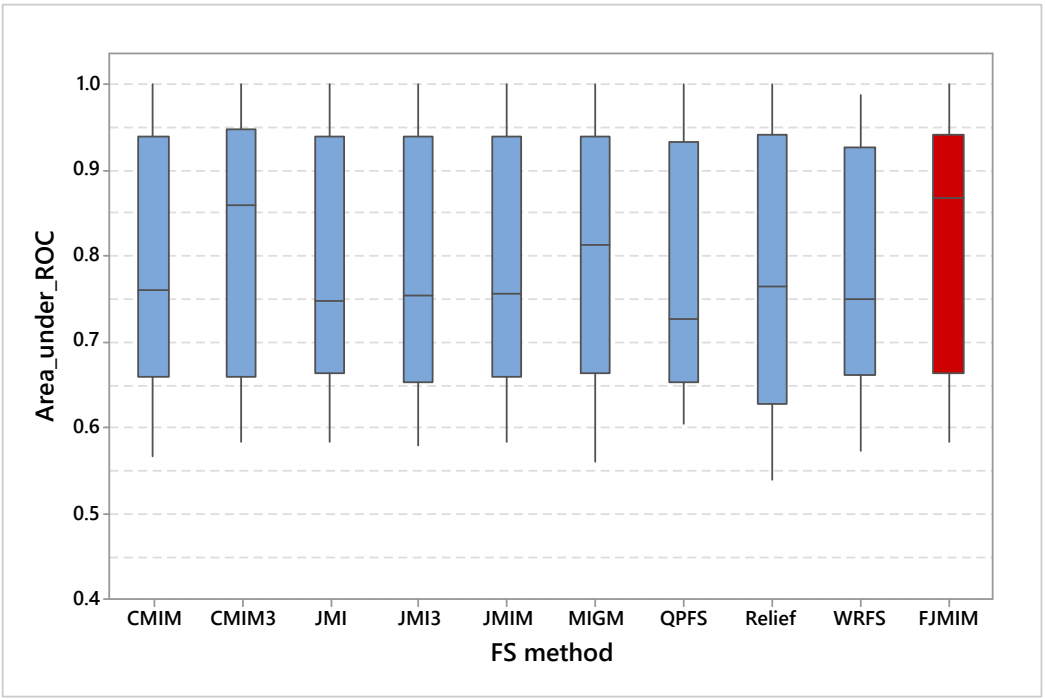


Figure 9. AUC distribution across all datasets.

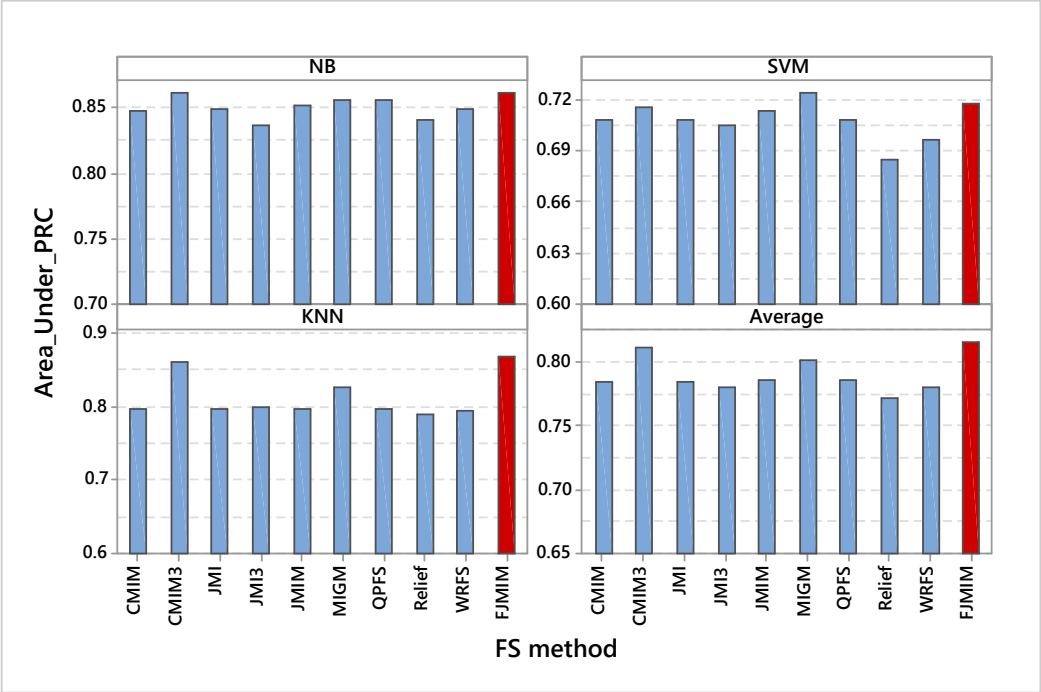


Figure 10. AUCPR result of NB, SVM, KNN and their average. FJMIM achieved the best result in all cases except SVM.

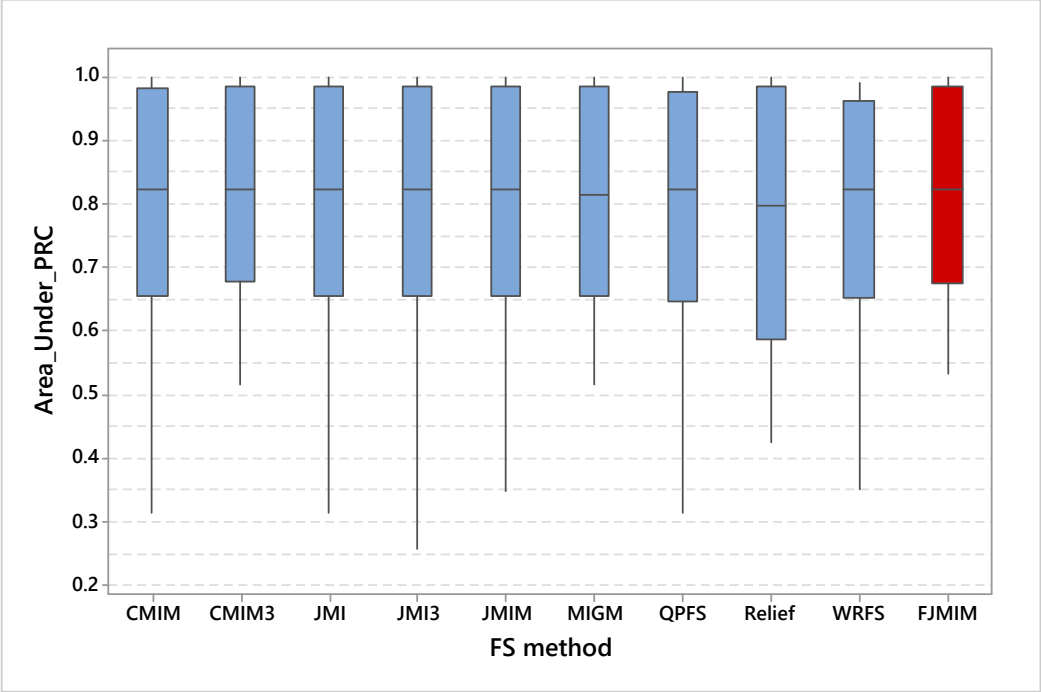
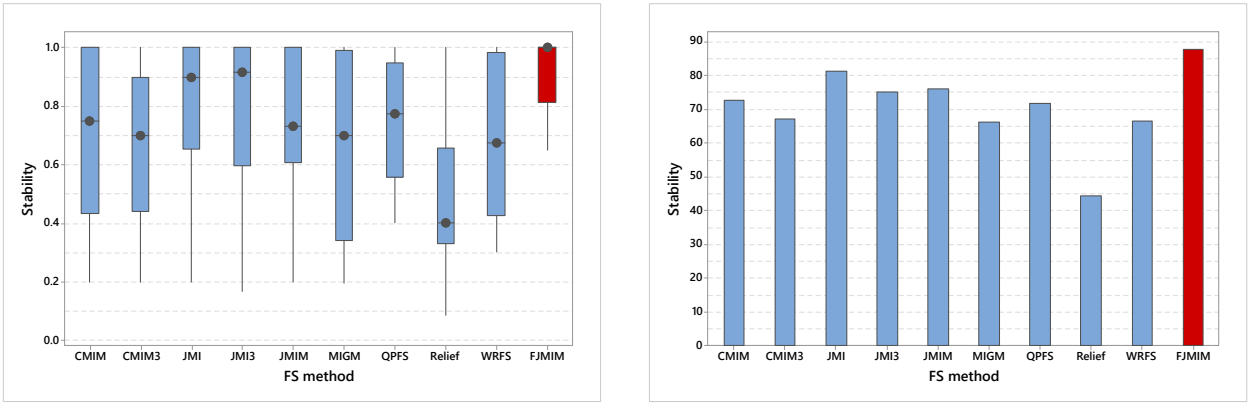


Figure 11. AUCPR distribution across all datasets.

5) AUCPR: The highest AUCPR was achieved by both FJMIM and CMIM3 using NB, MIGM using SVM, and FJMIM using KNN (Figure 10). On the other hand, Relief achieved the lowest AUCPR using all classifiers. According to the average of all classifiers, FJMIM achieved the best AUCPR by 81.6%, while Relief kept the lowest AUCPR by 77.2%. The proposed method outperformed other methods in the range from 0.3 to 4.4%. The second-best AUCPR was achieved by CMIM3, followed by MIGM, both QPFS and JMIM, both CMIM and JMI, and both JMI3 and WRFS. Figure 11 shows the distribution of AUCPR across all datasets. It's obvious that CMIM achieved the highest median, lower, and upper quartiles, while FJMIM achieved the highest median, upper quartile.

5.2. Feature stability

Figure 12(a) shows the stability of the used FS methods on all datasets. It's obvious that FJMIM is more consistent and stable compared to all other methods. Figure 12(b) confirms the stability of the compared method. FJMIM achieved the highest average stability by 87.8%. The proposed method outperformed other methods in the range from 6.6 to 43%. JMI achieved the second-best position by 81.2%, while Relief achieved the lowest stability by 44.3%.



(a) Stability distribution across all datasets

(b) The average stability of all methods

Figure 12. Stability result of all compared methods. FJMIM achieved the best result in all cases.

More detailed results are presented in the appendix section. According to previous results, It is obvious that FJMIM achieves the best results in most measures. This is expected because our proposed addresses the feature overestimation problem and handle the candidate feature problem well. Moreover, it avoids the discretization step. Another reason is the advantages of both inner and outer class information which FJMIM depends on it. This information helps the proposed method to be more robust toward the noise. On the other hand, the other compared methods are close to FJMIM than Relief that achieved the lowest result. This is because all compared method except Relief depends on mutual information as the proposed method to estimate the significant of features.

6. Conclusions

In this paper, we propose a new FS method, called, Fuzzy Joint Mutual Information Maximization (FJMIM). The proposed method depends on integrating an improved JMIM objective function with fuzzy concept. The benefits of our proposed method include: 1) The ability to deal directly with discrete and continuous features. 2) The suitability to handle any kind of relations between features such as linear and non-linear relation. 3) The ability to take the advantages of inner and outer class information. 4) The robustness toward the noise. 5) The ability to select the most significant feature subset and avoiding the undesirable features.

To confirm the effectiveness of FJMIM, 13 benchmark datasets have been used to evaluate the proposed method in the term of classification performance (accuracy, precision, F-measure, AUC, and AUCPR) and feature selection stability. According to nine conventional and state-of-the-art feature selection methods, the proposed method achieved promising improvement on feature selection in the terms of classification performance and stability.

In future work, we plan to extend the proposed method to cover multi-label classification problem. Moreover, we plan to study the effectiveness of imbalanced data on the proposed method.

Acknowledgements

This research has been supported by the National Natural Science Foundation (61572368).

Conflict of interests

The authors declare no conflict of interest.

References

1. L. T. Vinh, S. Lee, Y. Park, B. J. d'Auriol, A novel feature selection method based on normalized mutual information, *Appl. Intell.*, **37** (2012), 100–120.
2. J. R. Vergara, P. A. Estévez, A review of feature selection methods based on mutual information, *Neural Comput. Appl.*, **24** (2014), 175–186.
3. I. K. Fodor, *A survey of dimension reduction techniques*, Lawrence Livermore National Lab, CA (US), 2002.
4. H. X. Li, L. D. Xu, Feature space theory—a mathematical foundation for data mining, *Knowl. Based Syst.*, **14** (2001), 253–257.
5. R. Thawonmas, S. Abe, A novel approach to feature selection based on analysis of class regions, *IEEE Trans. Syst. Man Cybern. Syst.*, **27** (1997), 196–207.
6. Y. Saeys, I. Inza, P. Larrañaga, A review of feature selection techniques in bioinformatics, *Bioinformatics*, **23** (2007), 2507–2517.
7. I. Guyon, A. Elisseeff, An introduction to variable and feature selection, *J. Mach. Learn. Res.*, **3** (2003), 1157–1182.

8. M. Bennasar, Y. Hicks, R. Setchi, Feature selection using joint mutual information maximisation, *Expert Syst. Appl.*, **42** (2015), 8520–8532.
9. Q. Hu, D. Yu, Z. Xie, Information-preserving hybrid data reduction based on fuzzy-rough techniques, *Pattern Recognit. Lett.*, **27** (2006), 414–423.
10. C. Lazar, J. Taminau, S. Meganck, D. Steenhoff, A. Coletta, C. Molter, et al., A survey on filter techniques for feature selection in gene expression microarray analysis, *IEEE/ACM Trans. Comput. Biol. Bioinf.*, **9** (2012), 1106–1119.
11. G. Chandrashekar, F. Sahin, A survey on feature selection methods, *Comput. Electr. Eng.*, **40** (2014), 16–28.
12. O. A. Salem, L. Wang, Fuzzy mutual information feature selection based on representative samples, *Int. J. Software Innovation*, **6** (2018), 58–72.
13. D. Mo, S. H. Huang, Feature selection based on inference correlation, *Intell. Data Anal.*, **15** (2011), 375–398.
14. R. Steuer, J. Kurths, C. O. Daub, J. Weise, J. Selbig, The mutual information: detecting and evaluating dependencies between variables, *Bioinformatics*, **18** (2002), S231–S240.
15. J. Wang, J. M. Wei, Z. Yang, S. Q. Wang, Feature selection by maximizing independent classification information, *IEEE Trans. Knowl. Data Eng.*, **29** (2017), 828–841.
16. F. Macedo, M. R. Oliveira, A. Pacheco, R. Valadas, Theoretical foundations of forward feature selection methods based on mutual information, *Neurocomputing*, **325** (2019), 67–89.
17. D. Yu, S. An, Q. Hu, Fuzzy mutual information based min-redundancy and max-relevance heterogeneous feature selection, *Int. J. Comput. Intell. Syst.*, **4** (2011), 619–633.
18. J. Liang, K. Chin, C. Dang, R. C. Yam, A new method for measuring uncertainty and fuzziness in rough set theory, *Int. J. Gen. Syst.*, **31** (2002), 331–342.
19. Z. Li, P. Zhang, X. Ge, N. Xie, G. Zhang, C. F. Wen, Uncertainty measurement for a fuzzy relation information system, *IEEE Trans. Fuzzy Syst.*, **27** (2019), 2338–2352.
20. C. Wang, Y. Huang, M. Shao, D. Chen, Uncertainty measures for general fuzzy relations, *Fuzzy Sets Syst.*, **360** (2019), 82–96.
21. Y. Li, K. Qin, X. He, Some new approaches to constructing similarity measures, *Fuzzy Sets Syst.*, **234** (2014), 46–60.
22. G. Brown, A new perspective for information theoretic feature selection, *Artif. Intell. Stat.*, 2009, 49–56.
23. D. D. Lewis, *Feature selection and feature extract ion for text categorization*, Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, 1992, 23–26.
24. R. Battiti, Using mutual information for selecting features in supervised neural net learning, *IEEE Trans. Neural Netw. Learn. Syst.*, **5** (1994), 537–550.
25. H. Peng, F. Long, C. Ding, Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy, *IEEE Trans. Pattern Anal. Mach. Intell.*, **27** (2005), 1226–1238.

26. H. Yang, J. Moody, Feature selection based on joint mutual information, *Proc. Int. ICSC Symp. Adv. Intell. Data Anal.*, 1999, 22–25.
27. F. Fleuret, Fast binary feature selection with conditional mutual information, *J. Mach. Learn. Res.*, **5** (2004), 1531–1555.
28. P. E. Meyer, G. Bontempi, *On the use of variable complementarity for feature selection in cancer classification*, Workshops on applications of evolutionary computation, Springer, Berlin, Heidelberg, 2006, 91–102.
29. A. El Akadi, A. El Ouardighi, D. Aboutajdine, A powerful feature selection approach based on mutual information, *Int. J. Comput. Sci. Network Secur.*, **8** (2008), 116.
30. P. A. Estévez, M. Tesmer, C. A. Perez, J. M. Zurada, Normalized mutual information feature selection, *IEEE Trans. Neural Networks*, **20** (2009), 189–201.
31. N. Hoque, D. Bhattacharyya, J. K. Kalita, Mifs-nd: a mutual information-based feature selection method, *Expert Syst. Appl.*, **41** (2014), 6371–6385.
32. G. Herman, B. Zhang, Y. Wang, G. Ye, F. Chen, Mutual information-based method for selecting informative feature sets, *Pattern Recognit.*, **46** (2013), 3315–3327.
33. J. Y. Ching, A. K. Wong, K. C. C. Chan, Class-dependent discretization for inductive learning from continuous and mixed-mode data, *IEEE Trans. Pattern Anal. Mach. Intell.*, **17** (1995), 641–651.
34. Q. Shen, R. Jensen, Selecting informative features with fuzzy-rough sets and its application for complex systems monitoring, *Pattern Recognit.*, **37** (2004), 1351–1363.
35. J. Zhao, Z. Zhang, C. Han, Z. Zhou, Complement information entropy for uncertainty measure in fuzzy rough set and its applications, *Soft Comput.*, **19** (2015), 1997–2010.
36. H.-M. Lee, C.-M. Chen, J.-M. Chen, Y.-L. Jou, An efficient fuzzy classifier with feature selection based on fuzzy entropy, *IEEE Trans. Syst. Man Cybern. Syst.*, **31** (2001), 426–432.
37. I. Rodriguez-Lujan, R. Huerta, C. Elkan, C. S. Cruz, Quadratic programming feature selection, *J. Mach. Learn. Res.*, **11** (2010), 1491–1516.
38. K. Kira, L. A. Rendell, *The feature selection problem: Traditional methods and a new algorithm*, Aaai, **2** (1992), 129–134.
39. K. Sechidis, L. Azzimonti, A. Pocock, G. Corani, J. Weatherall, G. Brown, Efficient feature selection using shrinkage estimators, *Mach. Learn.*, **108** (2019), 1261–1286.
40. X. Wang, B. Guo, Y. Shen, C. Zhou, X. Duan, Input feature selection method based on feature set equivalence and mutual information gain maximization, *IEEE Access*, **7** (2019), 151525–151538.
41. P. Zhang, W. Gao, G. Liu, Feature selection considering weighted relevancy, *Appl. Intell.*, 1–11.
42. S. Garcia, J. Luengo, J. A. Sáez, V. Lopez, F. Herrera, A survey of discretization techniques: Taxonomy and empirical analysis in supervised learning, *IEEE Trans. Knowl. Data Eng.*, **25** (2012), 734–750.
43. A. Tharwat, Classification assessment methods, *Appl. Comput. Inform.*, 2020.
44. M. Allahyari, S. Pouriyeh, M. Assefi, S. Safaei, E. D. Trippe, J. B. Gutierrez, et al., A brief survey of text mining: Classification, clustering and extraction techniques, preprint, `arXiv:1707.02919`.

45. R. Kohavi, *A study of cross-validation and bootstrap for accuracy estimation and model selection*, Ijcai, **14** (1995),1137–1145.

46. S. Nogueira, G. Brown, *Measuring the stability of feature selection*, Joint European conference on machine learning and knowledge discovery in databases, Springer, Cham, 2016, 442–457.

47. Y. S. Tsai, U. C. Yang, I. F. Chung, C. D. Huang, *A comparison of mutual and fuzzy-mutual information-based feature selection strategies*, 2013 IEEE international conference on fuzzy systems (FUZZ-IEEE), IEEE, 2013, 1–6.

48. L. I. Kuncheva, *A stability index for feature selection*, Artificial intelligence and applications, 2007, 421–427.

49. D. Dua, C. Graff, UCI machine learning repository, 2017. Available from: http://archive.ics.uci.edu/ml.

Appendix

Tables A1–A6 show some numerical results of Figures 3(b), 4, 6, 8, 10 and 12(b), while Figures A1–A4 show the statistical results (mean $\pm$ standard deviation) of some datasets (DRD, Sonar, Wilt and Zoo) by precision, F-measure, AUC and AUCPR, respectively.

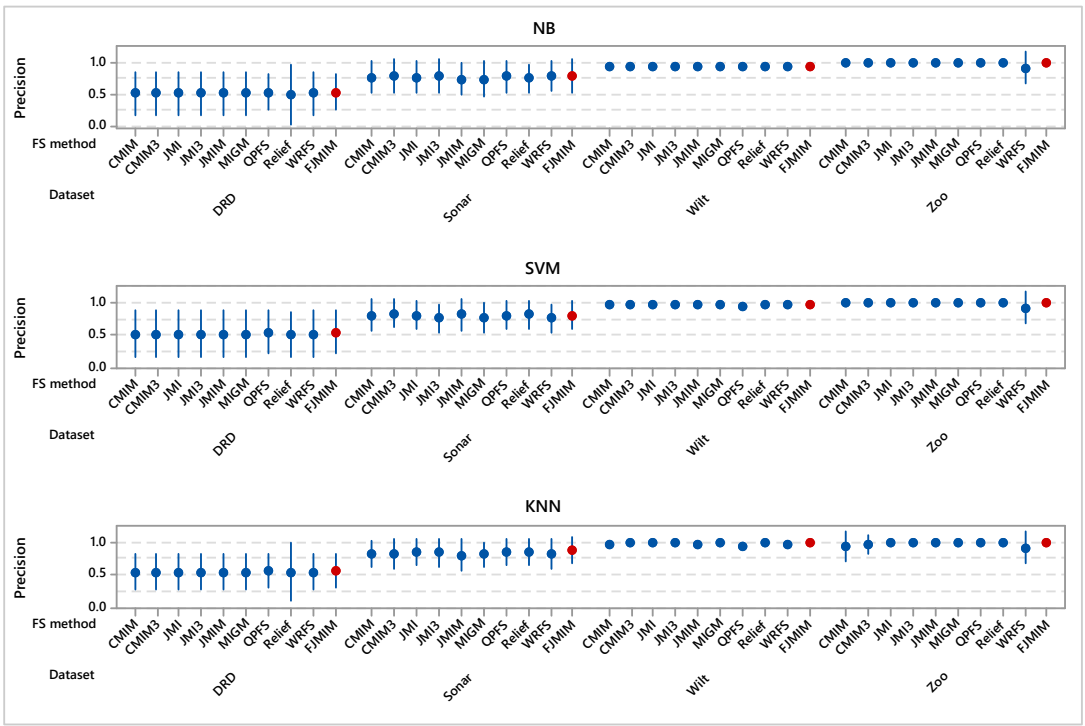


Figure A1. The precision results of four datasets (DRD, Sonar, Wilt and Zoo) on the used classifiers (NB, SVM and KNN).

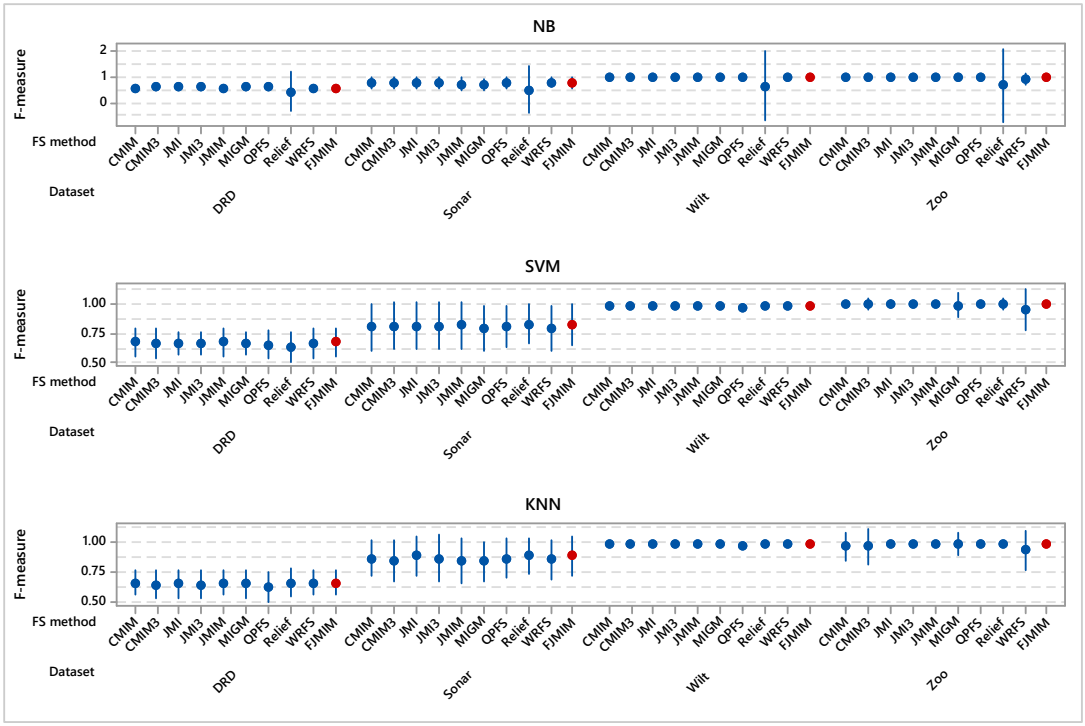


Figure A2. The F-measure results of four datasets (DRD, Sonar, Wilt and Zoo) on the used classifiers (NB, SVM and KNN).

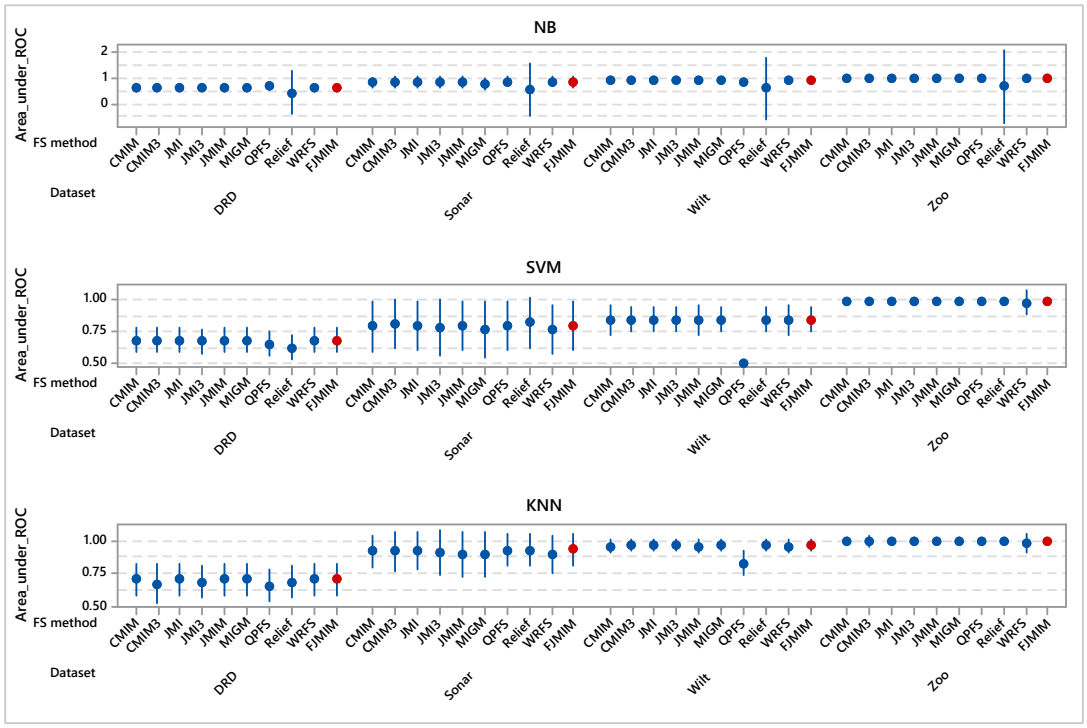


Figure A3. The AUC results of four datasets (DRD, Sonar, Wilt and Zoo) on the used classifiers (NB, SVM and KNN).

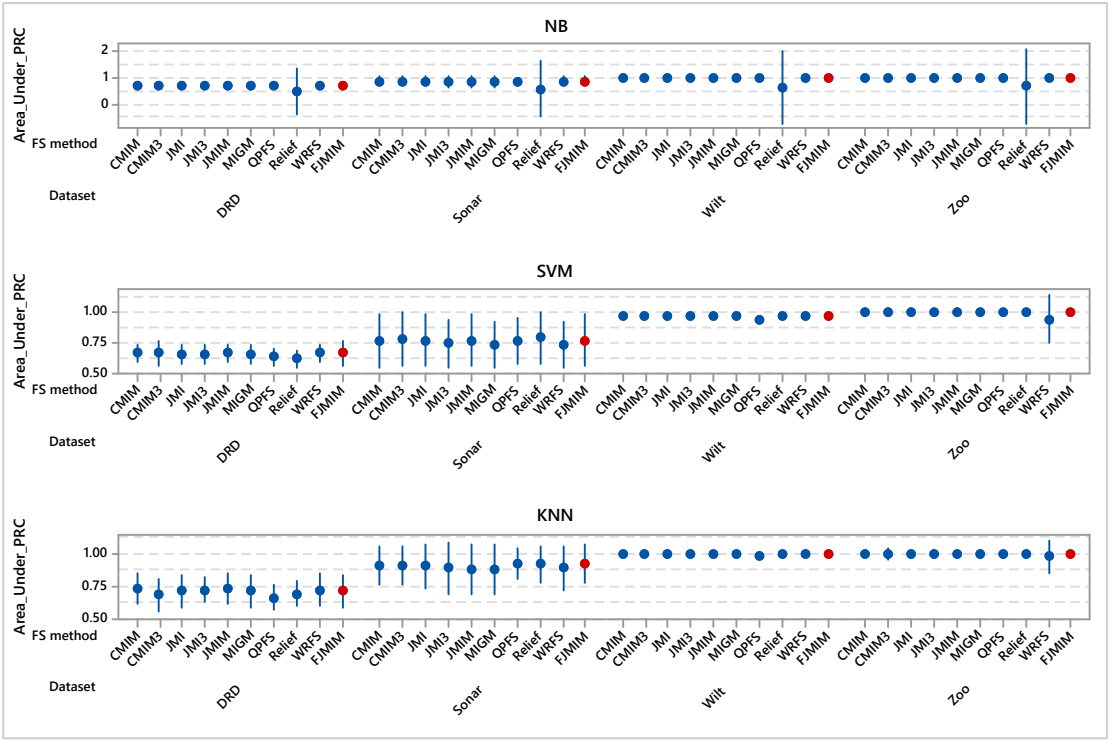


Figure A4. The AUCPR results of four datasets (DRD, Sonar, Wilt and Zoo) on the used classifiers (NB, SVM and KNN).

Table A1. Average accuracy of compared FS methods.

FS method	Average
CMIM	78.9
CMIM3	78.6
JMI	78.9
JMI3	79.1
JMIM	78.6
MIGM	78.4
QPFS	79.2
Relief	76.9
WRFS	78.1
FJMIM	79.8

Table A2. Precision results according to the used classifiers and their average.

FS method	NB	SVM	KNN	Average
CMIM	0.702	0.641	0.704	0.682
CMIM3	0.742	0.641	0.705	0.696
JMI	0.705	0.640	0.713	0.686
JMI3	0.712	0.636	0.723	0.690
JMIM	0.701	0.642	0.707	0.683
MIGM	0.701	0.637	0.708	0.682
QPFS	0.712	0.636	0.711	0.686
Relief	0.621	0.612	0.661	0.631
WRFS	0.698	0.632	0.708	0.679
FJMIM	0.748	0.645	0.743	0.712

Table A3. F-measure results according to the used classifiers and their average.

FS method	NB	SVM	KNN	Average
CMIM	0.722	0.712	0.818	0.750
CMIM3	0.795	0.712	0.752	0.753
JMI	0.724	0.712	0.823	0.753
JMI3	0.729	0.712	0.834	0.759
JMIM	0.719	0.714	0.821	0.751
MIGM	0.722	0.710	0.818	0.750
QPFS	0.732	0.712	0.825	0.756
Relief	0.632	0.708	0.688	0.676
WRFS	0.716	0.701	0.826	0.748
FJMIM	0.798	0.715	0.840	0.784

Table A4. AUC results according to the used classifiers and their average.

FS method	NB	SVM	KNN	Average
CMIM	0.835	0.733	0.812	0.793
CMIM3	0.836	0.748	0.844	0.809
JMI	0.837	0.734	0.815	0.795
JMI3	0.831	0.729	0.805	0.788
JMIM	0.836	0.742	0.812	0.796
MIGM	0.832	0.754	0.807	0.798
QPFS	0.838	0.710	0.807	0.785
Relief	0.804	0.707	0.801	0.771
WRFS	0.831	0.732	0.806	0.789
FJMIM	0.839	0.748	0.852	0.813

Table A5. AUCPR results according to the used classifiers and their average.

FS method	NB	SVM	KNN	Average
CMIM	0.848	0.708	0.798	0.785
CMIM3	0.862	0.715	0.862	0.813
JMI	0.850	0.708	0.797	0.785
JMI3	0.837	0.705	0.800	0.781
JMIM	0.852	0.713	0.797	0.787
MIGM	0.857	0.724	0.827	0.803
QPFS	0.857	0.708	0.797	0.787
Relief	0.842	0.685	0.791	0.772
WRFS	0.850	0.697	0.795	0.781
FJMIM	0.862	0.718	0.869	0.816

Table A6. Average stability of compared FS methods.

FS method	Stability
CMIM	72.6
CMIM3	67.3
JMI	81.2
JMI3	75.2
JMIM	75.9
MIGM	66.2
QPFS	71.7
Relief	44.3
WRFS	66.6
FJMIM	87.8