Fuzzy Information Measures Feature Selection Using Descriptive Statistics Data

Omar A. M. Salem[1,2], Haowen Liu[1(✉)], Feng Liu[1(✉)], Yi-Ping Phoebe Chen[3], and Xi Chen[1(✉)]

[1] School of Computer Science, Wuhan University, Wuhan 430072, China
{omarsalem,hwenliu,fliuwhu,robertcx}@whu.edu.cn
[2] Faculty of Computers and Informatics, Suez Canal University, Ismailia 41522, Egypt
omarsalem@ci.suez.edu.eg
[3] Department of Computer Science and Information Technology, La Trobe University, Melbourne 3086, Australia
phoebe.chen@latrobe.edu.au

Abstract. Feature selection (FS) has proven its importance as a preprocessing for improving classification performance. The success of FS methods depends on extracting all the possible relations among features to estimate their informative amount well. Fuzzy information measures are powerful solutions that extract the different feature relations without information loss. However, estimating fuzzy information measures consumes high resources such as space and time. To reduce the high cost of these resources, this paper proposes a novel method to generate FS based on fuzzy information measures using descriptive statistics data (DS) instead of the original data (OD). The main assumption behind this is that the descriptive statistics of features can hold the same relations as the original features. Over 15 benchmark datasets, the effectiveness of using DS has been evaluated on five FS methods according to the classification performance and feature selection cost.

Keywords: Feature selection · Fuzzy information measures · Fuzzy sets · Descriptive statistics · Classification systems

1 Introduction

Nowadays, classification systems can be founded in many real-world problems of different domains such as medical, software engineering, and industrial domain [3,10]. In real-world problems, classification data may contain a lot of features, but not all the features are significant [22]. The bad effect of irrelevant and redundant features reduces the classification performance and increases the computational cost of classification systems [21]. FS is an effective preprocessing on classification data to select the most informative feature subset by keeping only the relevant features and filtering out the undesirable features [2].

G. Memmi et al. (Eds.): KSEM 2022, LNAI 13370, pp. 77–90, 2022.
https://doi.org/10.1007/978-3-031-10989-8_7

The existing FS methods can be defined as one of three types [1,6]: filter, wrapper, and embedded. Our study focus on the filter type due to its advantage over other types as simplicity of usage, efficiency in the computational cost, and independence of classifiers [1,15].

Information measures are popular and widely used in the filter type [21]. Estimating these measures requires discretizing the continuous features with the risk of information loss [23]. To avoid this risk, fuzzy information measures have been introduced as an extension of information measures, by mapping each feature into a fuzzy relation matrix (FRM). The matrix size expands with increasing the length of features. Thus, estimating the fuzzy information measures consumes high computational cost in the space and time resources [18]. To reduce the high cost of these resources, this paper proposes a novel method to generate FS based on fuzzy information measures using descriptive statistics data (DS) instead of the original data (OD). The main assumption behind this is that the descriptive statistics of features can hold the same relations as the original features.

In this paper, the remaining sections are designed as follows: Sect. 2 introduces the related work. Section 3 presents the proposed method. The design of the experiment is described in Sect. 4 while the results are analyzed in Sect. 5. Finally, Sect. 6 introduces the conclusion of this paper.

2 Related Work

According to the structure of FRM, FS methods based on fuzzy information measures can be divided into two categories: FS based on feature-vector relationship and FS based on feature-feature relationship. However, both categories require high computational cost for mapping the original features into a FRM to avoid the discretization process.

In the category of FS based on feature-vector, Luukka et al. introduced FS method based on fuzzy entropy, called FES, to estimate the informative amount of each feature [12]. FES depends on the relationship between the feature and its ideal vector to map the feature into a FRM. Ideal vector is a user-defined set of samples that represents the class information as possible [12,18]. The highest informative feature (lowest entropy) is suggested for selection while the lowest informative feature (highest entropy) is suggested for denying. An improved version of FES is proposed, called FSAE [11]. FSAE depends on an additional scaling factor to consider the distance among ideal vectors with the aim to adjust the informative level of each ideal vector. Shen et al. [19] conducted a comparison among the different components of FES method to study the effect of the combination among the different components. The main limitation of the previous methods is that no consideration for important feature relations such as redundancy and complementarity.

In the second category, Hu et al. proposed FS method based on fuzzy entropy to deal with the heterogeneous data [8]. In [7], Hu et al. used a positive region of data to improve the original method. However, these methods still suffer from denying important feature relations such as redundancy and complementarity.

To overcome this drawback, Yu et al. proposed FS methods based on fuzzy mutual information, called FMI-mRMR [23]. In [20], Tsai et al. conducted a detailed comparison between FS methods based on mutual and fuzzy mutual information. The experimental results confirm the outperformance of FS method based on fuzzy mutual information in terms of feature stability and classification accuracy. Salem et al., in [16], proposed an ensemble FS method, called FFS-RRD, which depends on fuzzy information and fuzzy rough measures to extract the different feature relations. In [17], Salem et al. proposed a new FS method based on fuzzy joint mutual information, called FJMI, to extract the different relations based on the joint discriminative ability, in contrast to the traditional methods which depend on the individual discriminative ability.

3 Proposed Method

Most of the current FS methods depend on the original features. In this paper, we propose using descriptive statistics to summarize the feature information and reduce the computational cost of FS methods.

3.1 Fuzzy Relation Matrix Based on the Original Data

In the following, we illustrate the FRM structure based on the original data according to the methods of feature-feature relationship and feature-vector relationship.

Feature-Feature Relationship: Suppose $F = \{x_1, x_2, ..., x_m\}$ is a feature of m samples. The FRM between the feature and itself will be as follows:

$$M(F) = \begin{bmatrix} s_{1*1} & s_{1*2} & \cdots & s_{1*m} \\ s_{2*1} & \cdots & \cdots & s_{2*m} \\ \cdots & \cdots & \cdots & \cdots \\ s_{m*1} & s_{m*2} & \cdots & s_{m*m} \end{bmatrix} \tag{1}$$

where $s_{i*j} \in [0, 1]$ is the similarity degree between x_i and x_j, where $i, j \in \{1, 2, \ldots, m\}$.

Feature-Vector Relationship: Suppose $F = \{x_1, x_2, ..., x_m\}$ is a feature of m samples and $V = \{y_1, y_2, ..., y_t\}$ is an ideal vector of t samples. The FRM between the feature and its ideal vector will be as follows:

$$M(F) = \begin{bmatrix} s_{1*1} & s_{1*2} & \cdots & s_{1*t} \\ s_{2*1} & \cdots & \cdots & s_{2*t} \\ \cdots & \cdots & \cdots & \cdots \\ s_{m*1} & s_{m*2} & \cdots & s_{m*t} \end{bmatrix} \tag{2}$$

where $s_{i*j} \in [0, 1]$ is the similarity degree between x_i and y_j, where $i \in \{1, 2, \ldots, m\}$ and $j \in \{1, 2, \ldots, t\}$.

3.2 Descriptive Statistics

Descriptive statistics (DS) is a set of measures that describe the structure of data [4,9]. DS has two main types of measures: central tendency and dispersion (variation). Central tendency is a single measurement that describes the set of samples via their average, midpoint, and most frequently sample. Measures of dispersion (variation) describe how much the samples vary or are close to the central tendency. In this study, we use well-known statistics measures such as minimum (min), maximum (max), mean, median, mode, and standard deviation (Std). The basic definitions of the six measures are as follows:

– **Mean (or arithmetic mean)** is the average value of a set of samples.
– **Median** is the midpoint value of an ordered set of samples.
– **Mode** is the most frequently occurring sample in the set of samples.
– **Minimum (Min)** is the lowest value in a set of samples.
– **Maximum (Max)** is the highest value in a set of samples.
– **Standard deviation (Std)** is a spread measure that describes how much each sample varies or is close to the mean of the set of samples.

3.3 FS Based on Descriptive Statistics

Traditional FS methods of fuzzy information measures depend on the FRM to represent the feature structure as possible. However, generating the FRM is expensive in the space and time cost. To overcome the cost limitations, we suggest generating the FRM by the descriptive statistics data instead of original data. Based on DS, the values of six statistics measures can represent the samples of feature with respect to the class label. In this way, we can reduce the size of FRM as well as cost. Figure 1 shows the main process of FS based on DS. Firstly, we calculate the descriptive statistics of the original data. Then, we apply the FS method on DS. The indexes of the selected features are used to return the selected features from the original data.

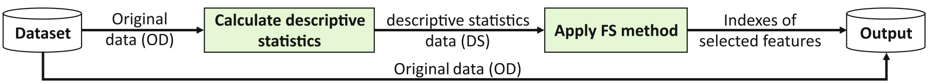

Fig. 1. The main process of FS based on descriptive statistics data.

The main procedure for generating a new dataset based on DS is described in Algorithm 1. The input of the algorithm is a dataset D, which consists of a set of features F and class label C. Firstly, we initialize the output dataset $Newdata$ as an empty list (line 1). Then, we divide the dataset into subsets of data $Fsubset$, according to the class label (lines 2–3). Line 4 defines an empty list $Fnew$ to store the descriptive statistics for each feature f_i in $Fsubset$. The descriptive statistics of f_i are calculated, stored in $Dstat$, and added to

$Fnew$ list as shown in lines (5–9). After that, the class label is added to $Fnew$ (line 10). Each $Fnew$ of class c_i is appended to the final output $Newdata$. Finally, a $Newdata$ is returned in line 13.

Algorithm 1. Dataset based on descriptive statistics

Input: A dataset $D = \langle F \cup C \rangle$, where F is a set of features and C is the class label.
Output: new data of descriptive statistics $Newdata$.
1: $Newdata \leftarrow [\ ]$
2: **for each** $c_i \in C$ **do**
3: $\quad Fsubset \leftarrow F(c_i)$
4: $\quad Fnew \leftarrow [\ ]$
5: $\quad$ **for each** $f_j \in Fsubset$ **do**
6: $\quad\quad Dstat \leftarrow [\ ]$
7: $\quad\quad Dstat[] \leftarrow [mean(f_j), median(f_j), mode(f_j), min(f_j), max(f_j), std(f_j)]$
8: $\quad\quad Fnew \leftarrow Fnew \cup \{Dstatc\}$
9: $\quad$ **end for**
10: $\quad Fnew \leftarrow Fnew \cup \{c_i * ones(6, 1)\}$
11: $\quad Newdata \leftarrow append(Newdata, Fnew)$
12: **end for**
13: **return** $Newdata$

4 Experimental Design

The main phases of the experimental framework (data preparation, feature selection, and evaluation) are designed as shown in Fig. 2.

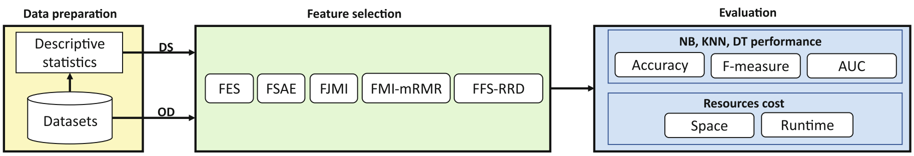

Fig. 2. The main phases of the experimental framework.

4.1 Data Preparation

To justify the effectiveness and efficiency of using DS, the experiment was conducted on 15 benchmark datasets collected from[1,2]. Table 1 shows the main properties of the used datasets. In this phase, the output is the original data (OD) and a new data of descriptive statistics (DS).

1 https://archive.ics.uci.edu/ml/datasets.php.
2 https://github.com/klainfo/NASADefectDataset.

Table 1. The main properties of the used datasets.

Dataset	Abbreviation	#Features	#instances	#classes
CM1	CM1	37	327	2
Credit approval	CAP	15	690	2
Glioma	GLO	4434	50	4
JMI	JMI	21	7720	2
KC1	KC1	21	1162	2
KC3	KC3	39	194	2
MC1	MC1	38	1952	2
MC2	MC2	39	124	2
MW1	MW1	37	250	2
SPECTF Heart	NHE	44	267	2
SPECT Heart	SHE	22	267	2
DNA	DNA	180	2000	3
Multiple features	MFE	649	2000	10
Ozone level detection	OLD	72	1848	2
seismic-bumps	SBU	18	2584	2

4.2 Feature Selection

To confirm the effectiveness of using DS, five FS methods (with 50% threshold of ranked features) have been used in the experimental comparison. FS methods is divided into two categories: feature-vector methods (FES [12] and FSAE [11]) with time complexity $O(mtn)$ and feature-feature methods (FJMI [17], FMI-mRMR [23], and FFS-RRD [16]) with time complexity $O(dnm^2)$, where m is the number of samples in the feature, t is the number of samples in the ideal vector, n is the total number of features, and d is the number of selected information.

4.3 Evaluation

The evaluation of our experiment depends on two parts: classification performance and feature selection cost.

Classification Performance: In the experiment, three well-known classifiers are used to verify the improvement of classification performance as Naive Bayes (NB) [5], k-Nearest Neighbors (KNN, K = 3) [13], and Decision Tree (DT) [13]. The main measures of classification performance are:

1- Accuracy: is the percentage of the correctly predicted instances.
2- F-measure: is the harmonic average of the classification precision and recall.

3- Area under the ROC curve (AUC): AUC is the size of the area under the ROC curve. ROC is a curve graph that represents the relation between the true positive rate and the false positive rate.

Feature Selection Cost: In this paper, the experiments were conducted in a computer system with Ryzen 7 4800H (2.9 GHz) CPU and 16 GB RAM.

4- Space cost: the space cost is defined by the size of the FRM. The matrix size of each feature with OD is $m * t$ for feature-vector methods while m^2 for feature-feature methods, where m is the number of samples in the feature and t is the number of samples in the ideal vector. For DS, the matrix size of each feature is $(6 * h) * t$ for feature-vector methods while $(6 * h)^2$ for feature-feature methods, where h is the number of classes. The reduced percentage of the FRM was also computed to show the reduction size that DS achieved compared to OD as [14]:

$$MR(\%) = 100 - \frac{w1}{w2} * 100 \tag{3}$$

where $w1$ is the relation matrix size of the feature with DS and $w2$ is the relation matrix size with OD.
5- Runtime cost: the execution time of the FS methods represents the runtime cost. The reduced percentage of time was also computed to show the reduction time that DS achieved compared to OD as [14]:

$$TR(\%) = 100 - \frac{r1}{r2} * 100 \tag{4}$$

where $r1$ is the runtime of DS and $r2$ is the runtime of OD.

5 Results and Analysis

5.1 Accuracy

The accuracy results of NB obtained by the different FS methods are shown in Table 2. Using DS improved the average accuracy of all FS methods with OD. FES with DS (for simplicity, FES (DS)) improved FES (OD) by 0.2%. Similarly, DS improved the average accuracy of FSAE, FJMI, FMI-mRMR, and FFS-RRD with OD by 0.5%, 5.9%, 0.3%, and 3.2%, respectively.

Table 3 shows the accuracy results of KNN obtained by the different FS methods. Among five methods, DS improved the average accuracy of three methods FJMI, FMI-mRMR, and FFS-RRD with OD by 2.8%, 1.2%, and 0.2%, respectively. For methods of FES and FSAE, OS improved the average accuracy of the used methods with DS by 0.6% and 0.3%, respectively.

The accuracy results of DT obtained by the different FS methods are shown in Table 4. DS improved the average accuracy of FES, FSAE, FJMI, FMI-mRMR, and FFS-RRD with OD by 0.3%, 0.1%, 2.3%, 0.4%, and 1.9%, respectively.

5.2 F-measure

Figure 3 shows the F-measure results of the three classifiers obtained by the FS methods. According to NB, DS improved the average F-measure of FES, FSAE, FJMI, FMI-mRMR, and FFS-RRD by 1.1%, 0.3%, 5%, 0.7%, and 1%. Using KNN, FES(OD), FSAE(OD), and FFS-RRD(OD) have more average F-measure than the same methods with DS by 0.8%, 0.1%, and 0.9%, respectively. FJMI(DS) and FMI-mRMR(DS) outperformed the original methods by 3.8% and 1.4%, respectively. For DT, DS improved the average F-measure of all used FS methods except FES. In FES, using OD has more F-measure by 0.5%. The remaining methods FSAE(DS), FJMI(DS), FMI-mRMR(DS), and FFS-RRD(DS) achieved more average F-measure compared to the original methods by 0.6%, 1.5%, 0.5%, and 1.9%, respectively.

Table 2. Classification accuracy derived by NB classifier among five FS methods using OD and DS. On average, DS outperformed OD in all FS methods.

Dataset	FES		FSAE		FJMI		FMI-mRMR		FFS-RRD	
	OD	DS	OD	DS	OD	DS	OD	DS	OD	DS
CM1	67.4	65.8	67.1	65.7	77.9	84.3	70.8	70.5	70.9	79.5
CAP	85	85	83.4	83.6	76.8	86.1	77	84.5	68.4	76.9
GLO	62.4	62	63.2	65.2	73.6	84.6	66.2	67.8	65.4	67.8
JMI	67.3	65.9	67.3	65.9	67	67.7	68.9	69.9	68.4	73.3
KC1	65.4	66	65.4	66.3	64.8	74.7	66.4	67.3	66.3	69.7
KC3	63	66.2	63	67.1	70.7	83.5	68.4	67	67.3	75.6
MC1	74.3	78.5	74.3	79.1	87	92.8	81.9	83.3	83	90.7
MC2	65.4	62.5	66	62.5	69.6	68.8	67.4	66.8	68.2	69.6
MW1	71.5	72.5	71.3	74	82	85.3	76.1	77	76.7	82.9
NHE	71.5	69.6	80.7	80.2	80.5	77.8	78.6	78.6	75.3	76.3
SHE	73.7	75.3	75.4	76.3	75.8	79.8	76.1	78.4	76.7	73.2
DNA	95.1	95.1	95.3	95.2	73.8	78.8	94.5	91.3	90.7	90.7
MFE	95.3	96.2	95.2	96	81.1	94	95.6	95.8	90.2	90.2
OLD	76.8	76.1	79	76.2	74.6	83.3	80.7	77.7	75.7	72
SBU	88.9	90	88.9	90.1	90.4	93.4	90.7	88.3	90	92.6
Average	74.9	75.1	75.7	76.2	76.4	82.3	77.3	77.6	75.5	78.7

Table 3. Classification accuracy derived by KNN classifier among five FS methods using OD and DS. On average, DS outperformed OD in three of five FS methods.

Dataset	FES		FSAE		FJMI		FMI-mRMR		FFS-RRD	
	OD	DS	OD	DS	OD	DS	OD	DS	OD	DS
CM1	85.6	82.7	83.9	82.5	84.2	82.9	82.3	82.7	82.8	84.3
CAP	85.4	85.4	87.5	85	76.3	85.5	73.8	82.4	63.5	74.5
GLO	77.8	72.8	78.4	74.4	78.4	84.6	77.8	79.8	78.2	74
JMI	76.5	77.2	76.5	77.2	76.3	76.8	75.8	76.1	76.1	77.9
KC1	73.5	74.1	73.5	73.7	75	76.9	74.7	74.2	75.4	72.8
KC3	82.1	81.6	82.1	80.9	77.5	79.9	79	81.6	78.3	84.4
MC1	98	97.9	98	97.9	98.1	97.9	97.9	97.9	98.2	98
MC2	70	65	68.2	67.7	58.7	64.7	71.4	70.1	70.8	65
MW1	88.6	89.1	88.2	88.4	90.4	89	90.1	90.2	89.9	89.3
NHE	78.7	77.9	79.9	79.3	79.9	81.1	79.5	79.5	78.5	79.4
SHE	75	78.2	72.1	75.6	75	76.7	72.6	78	76.5	72
DNA	69.8	69.8	73.4	73.7	73.8	78.2	73.7	74.3	60.5	60.5
MFE	97.4	97.6	97.4	97.6	86.9	96.5	97.9	97.6	95.8	95.8
OLD	96.1	96.5	96	96.4	96.1	96.4	96.2	96.5	96.2	96.8
SBU	92.2	92.4	92.2	92.6	92	93.4	93.3	93.2	92.4	92.2
Average	83.1	82.5	83.2	82.9	81.2	84.0	82.4	83.6	80.9	81.1

Table 4. Classification accuracy derived by DT classifier among five FS methods using OD and DS. On average, DS outperformed OD in all FS methods.

Dataset	FES		FSAE		FJMI		FMI-mRMR		FFS-RRD	
	OD	DS	OD	DS	OD	DS	OD	DS	OD	DS
CM1	84	82.4	84.3	82.5	87	87.2	81.8	82.8	82.3	86.9
CAP	85	85	85.4	87.2	75.4	84.7	75.7	84.4	67.9	75.8
GLO	55.2	54.8	47.6	49.4	56	56.6	43.4	41	44	49.2
JMI	78.5	79	78.5	79	78.3	78.8	78.7	78.6	78.6	79.2
KC1	76.9	76	76.9	75.1	74.6	76.9	76.1	75.6	75.5	74.1
KC3	81.8	82	81.8	81.7	79.6	81.4	82.4	79.9	78.8	81.6
MC1	98.2	98.2	98.2	98.2	98.2	98.2	98.2	98.2	98.2	98.2
MC2	64.7	66.6	67.8	71.4	67.8	70.6	65.6	68.8	68.3	71.1
MW1	88.4	88.6	90.4	88.7	89.2	87.9	88.8	89.4	88.6	89.8
NHE	79.4	79.4	80.8	78.6	77.5	78.5	79.1	78.9	76.7	79.4
SHE	73.3	75.4	74.4	75.3	75.1	80.3	73.8	75	76	77.6
DNA	92.5	92.5	92.9	92.8	73.8	78.7	92.9	88.6	87.3	87.3
MFE	93.3	95	93.5	94.8	84.6	92	94.2	94.3	88.5	88.5
OLD	96.1	96.1	96.1	96	96.7	96.9	95.9	96.2	95.7	96.9
SBU	93.4	93.4	93.4	93.4	93.4	93.4	93.4	93.4	93.4	93.4
Average	82.7	83.0	82.8	82.9	80.5	82.8	81.3	81.7	80.0	81.9

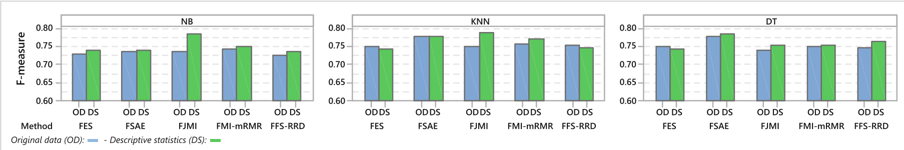

Fig. 3. Classification F-measure derived by the three classifiers among five FS methods using OD and DS. On average, DS outperformed OD in most FS methods using NB and DT. For KNN, OD outperformed DS in three FS methods

5.3 AUC

Figure 4 shows the AUC results of the used classifiers obtained by the FS methods. In FES, FS methods outperformed on NB with DS by 0.5% while outperformed on DT with OD by 0.2%. FES achieved the same result with DS and OD. Respectively, methods of FSAE, FJMI, and FMI-mRMR have been improved with DS by 1.5%, 1.9%, and 0.1% using NB, 0.1%, 4.1%, and 1.5% using KNN, and 1.9%, 2.8%, and 0.8% using DT. In FFS-RRD, the AUC was better with OD by 0.6%, 1.1%, and 0.9% using NB, KNN, and DT, respectively.

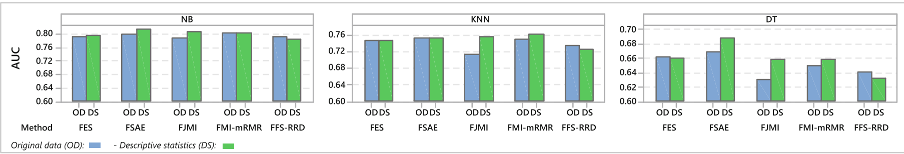

Fig. 4. Classification AUC derived by the three classifiers among five FS methods using OD and DS. On average, DS outperformed OD in most FS methods using.

5.4 Space Cost

Table 5 reports the relation matrix size of the feature in each dataset with OD and DS. It also shows the reduction percentage of matrix size (MR) induced by DS. It is obvious that FS methods with DS have a smaller matrix size than the same methods with OD. The reduction range induced by DS is from 52% to 99.84% using FES and FSAE while from 76.96% to around 100% using the remaining methods.

5.5 Runtime Cost

Table 6 reports the runtime efficiency on the FS methods with OD and DS. It also shows the reduction percentage of time (TR) induced by DS. It is obvious that FS methods with DS have a smaller runtime than the same methods with OD.

Table 5. Comparison of space cost on the FS methods between OD and DS. DS has the best space cost on all datasets

Dataset	FES/FSAE		MR(%)	FJMI/FMI-mRMR/FFS-RRD		MR(%)
	OD	DS		OD	DS	
CM1	327*2	(6*2)*2	96.330	327^2	$(6*2)^2$	99.865
CAP	690*2	(6*2)*2	98.261	690^2	$(6*2)^2$	99.970
GLO	50*4	(6*4)*4	52.000	50^2	$(6*4)^2$	76.960
JMI	7720*2	(6*2)*2	99.845	7720^2	$(6*2)^2$	≈100.00
KC1	1162*2	(6*2)*2	98.967	1162^2	$(6*2)^2$	99.989
KC3	194*2	(6*2)*2	93.814	194^2	$(6*2)^2$	99.617
MC1	1952*2	(6*2)*2	99.385	1952^2	$(6*2)^2$	99.996
MC2	124*2	(6*2)*2	90.323	124^2	$(6*2)^2$	99.063
MW1	250*2	(6*2)*2	95.200	250^2	$(6*2)^2$	99.770
NHE	267*2	(6*2)*2	95.506	267^2	$(6*2)^2$	99.798
SHE	267*2	(6*2)*2	95.506	267^2	$(6*2)^2$	99.798
DNA	2000*3	(6*3)*3	99.100	2000^2	$(6*3)^2$	99.992
MFE	2000*10	(6*10)*10	97.000	2000^2	$(6*10)^2$	99.910
OLD	1848*2	(6*2)*2	99.351	1848^2	$(6*2)^2$	99.996
SBU	2584*2	(6*2)*2	99.536	2584^2	$(6*2)^2$	99.998
Average	-	-	94.008	-	-	98.315

The reduction range induced by DS is from 83.9% to 99.99% using FES, 9.65% to 89.37% using FSAE, 1.36% to 76.22% using FJMI, 63.19% to 99.99% using FMI-mRMR, and 62.03% to 99.99% using FFS-RRD.

Table 6. Comparison of runtime cost on the FS methods between OD and DS. DS has the best runtime efficiency on all datasets.

Dataset	FES		FSAE		FJMI		FMI-mRMR		FFS-RRD	
	OD	DS	OD	DS	OD	DS	OD	DS	OD	DS
CM1	0.428742	0.02439	0.009218	0.008329	0.0096	0.006563	2.621383	0.067203	6.553457	0.08205
TR (%)	94.311		9.649		31.634		97.436		98.748	
CAP	1.214707	0.008385	0.00357	0.00146	0.003839	0.002375	1.914782	0.008474	4.786956	0.020408
TR (%)	99.310		59.105		38.116		99.557		99.574	
GLO	35.54285	5.721867	0.160168	0.066505	0.391724	0.35369	853.6609	314.2646	560.9826	213.0115
TR (%)	83.901		58.478		9.709		63.186		62.029	
JMI	373.4975	0.003991	0.038568	0.007782	0.022178	0.008228	495.8916	0.04638	1487.675	0.046238
TR (%)	99.990		79.822		62.901		99.991		99.990	
KC1	6.141769	0.002594	0.007138	0.002764	0.004871	0.002331	12.98244	0.011987	6.491221	0.027279
TR (%)	99.958		61.275		52.135		99.908		99.580	
KC3	0.130475	0.006057	0.003131	0.001104	0.002799	0.002429	1.661687	0.00957	4.154218	0.024887
TR (%)	95.358		64.747		13.218		99.424		99.401	

(*continued*)

Table 6. (*continued*)

Dataset	FES		FSAE		FJMI		FMI-mRMR		FFS-RRD	
	OD	DS	OD	DS	OD	DS	OD	DS	OD	DS
MC1	34.43672	0.0105	0.016591	0.003294	0.010812	0.005009	102.7852	0.023823	308.3556	0.036135
TR (%)	99.970		80.149		53.670		99.977		99.988	
MC2	0.059361	0.007061	0.002442	0.001036	0.002404	0.002371	0.309565	0.008864	0.681043	0.020617
TR (%)	88.105		57.575		1.361		97.137		96.973	
MW1	0.190526	0.006891	0.003237	0.001205	0.003057	0.002395	1.11319	0.009134	2.449017	0.024198
TR (%)	96.383		62.765		21.657		99.179		99.012	
NHE	0.101347	0.005195	0.002379	0.00081	0.002311	0.001595	0.360315	0.004761	0.792692	0.010015
TR (%)	94.874		65.951		30.968		98.679		98.737	
SHE	1.944846	0.004936	0.004029	0.001445	0.004111	0.002732	2.03347	0.01201	4.880328	0.031219
TR (%)	99.746		64.127		33.543		99.409		99.360	
DNA	39.75004	0.033494	0.109036	0.01316	0.071351	0.022158	1905.678	0.363506	622.712	0.485526
TR (%)	99.916		87.930		68.946		99.981		99.922	
MFE	482.2562	2.755366	0.708246	0.075309	0.794471	0.188911	29434.15	23.29409	8302.465	255.5522
TR (%)	99.429		89.367		76.222		99.921		96.922	
OLD	169.4312	0.017707	0.023083	0.006456	0.021521	0.008837	380.5805	0.068522	65.62529	0.083135
TR (%)	99.990		72.033		58.940		99.982		99.873	
SBU	13.17526	0.002211	0.005435	0.00181	0.005612	0.002438	40.37545	0.011567	113.0513	0.016277
TR (%)	99.983		66.699		56.563		99.971		99.986	
Average	77.220	0.574	0.073	0.013	0.090	0.041	2215.742	22.547	766.110	31.298
TR (%)	96.749		65.311		40.639		96.916		96.673	

Overall, It is obvious that FS methods with DS achieved the best classification performance in most cases. It justifies that summarizing the feature information by DS helps to define the feature information better. Moreover, DS reduced the size of FRM on each feature. This is because DS maps the feature into a smaller size of samples. As a result, the cardinal value of the feature based on DS is usually less than the cardinal value of the feature based on OD. This is also the same reason of why FS methods with DS have a smaller runtime than the same methods with OD.

6 Conclusion

Fuzzy information measures are powerful solutions for developing effective FS methods. However, the estimation cost of these measures is relative to the size of input data where increasing the former depends on increasing the latter. In this paper, we have introduced a novel method to reduce the high cost of FS methods based on fuzzy information measures. To achieve that, we generated descriptive statistics data (DS) from the original data (OD) to reduce the input data of FS methods. Consequently, the cost of FS methods based on fuzzy information measures has been reduced. The effectiveness of using DS has been evaluated on five FS methods. The experimental results confirm reducing the cost of FS methods and improving the classification performance in most cases. In future work, we plan to extend our study to cover more DS measures with the aim to highlight the importance of using DS for enhancing the FS process.

Acknowledgement. This research has been supported by the National Natural Science Foundation of China (62172309).

References

1. Bolón-Canedo, V., Sánchez-Maroño, N., Alonso-Betanzos, A.: Feature selection for high-dimensional data. Progr. Artif. Intell. **5**(2), 65–75 (2016). https://doi.org/10.1007/s13748-015-0080-y
2. Bommert, A., Sun, X., Bischl, B., Rahnenführer, J., Lang, M.: Benchmark for filter methods for feature selection in high-dimensional classification data. Comput. Stat. Data Anal. **143**, 106839 (2020)
3. Cateni, S., Colla, V., Vannucci, M.: A method for resampling imbalanced datasets in binary classification tasks for real-world problems. Neurocomputing **135**, 32–41 (2014)
4. Cavallaro, M., Fidell, L.: Basic descriptive statistics: commonly encountered terms and examples. Am. J. EEG Technol. **34**(3), 138–152 (1994)
5. Cheng, J., Greiner, R.: Comparing bayesian network classifiers. arXiv preprint arXiv:1301.6684 (2013)
6. Guyon, I., Gunn, S., Nikravesh, M., Zadeh, L.A.: Feature Extraction: Foundations and Applications, vol. 207. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-35488-8
7. Hu, Q., Xie, Z., Yu, D.: Hybrid attribute reduction based on a novel fuzzy-rough model and information granulation. Pattern Recogn. **40**(12), 3509–3521 (2007)
8. Hu, Q., Yu, D., Xie, Z.: Information-preserving hybrid data reduction based on fuzzy-rough techniques. Pattern Recogn. Lett. **27**(5), 414–423 (2006)
9. Kaur, P., Stoltzfus, J., Yellapu, V., et al.: Descriptive statistics. Int. J. Acad. Medi. **4**(1), 60 (2018)
10. Laradji, I.H., Alshayeb, M., Ghouti, L.: Software defect prediction using ensemble learning on selected features. Inf. Softw. Technol. **58**, 388–402 (2015)
11. Lohrmann, C., Luukka, P., Jablonska-Sabuka, M., Kauranne, T.: A combination of fuzzy similarity measures and fuzzy entropy measures for supervised feature selection. Exp. Syst. Appl. **110**, 216–236 (2018)
12. Luukka, P.: Feature selection using fuzzy entropy measures with similarity classifier. Exp. Syst. Appl. **38**(4), 4600–4607 (2011)
13. Patrick, E.A., Fischer, F.P.: A generalized k-nearest neighbor rule. Inf. Control **16**(2), 128–152 (1970)
14. Raza, M.S., Qamar, U.: Feature selection using rough set-based direct dependency calculation by avoiding the positive region. Int. J. Approx. Reason. **92**, 175–197 (2018)
15. Saeys, Y., Inza, I., Larrañaga, P.: A review of feature selection techniques in bioinformatics. bioinformatics **23**(19), 2507–2517 (2007)
16. Salem, O.A., Liu, F., Chen, Y.P.P., Chen, X.: Ensemble fuzzy feature selection based on relevancy, redundancy, and dependency criteria. Entropy **22**(7), 757 (2020)
17. Salem, O.A., Liu, F., Chen, Y.P.P., Chen, X.: Feature selection and threshold method based on fuzzy joint mutual information. Int. J. Approx. Reason. **132**, 107–126 (2021)
18. Salem, O.A., Liu, F., Chen, Y.P.P., Hamed, A., Chen, X.: Fuzzy joint mutual information feature selection based on ideal vector. Exp. Syst. Appl. **193**, 116453 (2022)

19. Shen, Z., Chen, X., Garibaldi, J.: Performance optimization of a fuzzy entropy based feature selection and classification framework. In: 2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp. 1361–1367. IEEE (2018)
20. Tsai, Y.S., Yang, U.C., Chung, I.F., Huang, C.D.: A comparison of mutual and fuzzy-mutual information-based feature selection strategies. In: 2013 IEEE International Conference on, Fuzzy Systems (FUZZ), pp. 1–6. IEEE (2013)
21. Vergara, J.R., Estévez, P.A.: A review of feature selection methods based on mutual information. Neural Comput. Appl. **24**(1), 175–186 (2013). https://doi.org/10.1007/s00521-013-1368-0
22. Xue, B., Zhang, M., Browne, W.N., Yao, X.: A survey on evolutionary computation approaches to feature selection. IEEE Trans. Evol. Comput. **20**(4), 606–626 (2015)
23. Yu, D., An, S., Hu, Q.: Fuzzy mutual information based min-redundancy and max-relevance heterogeneous feature selection. Int. J. Comput. Intell. Syst. **4**(4), 619–633 (2011)