





Information Theory-Based Feature Selection: Minimum Distribution Similarity with Removed Redundancy

Yu Zhang , Zhuoyi Lin, and Chee Keong Kwoh 

School of Computer Science and Engineering, Nanyang Technological
University, 50 Nanyang Avenue, Singapore 639798, Singapore
{YU007, ZHUOYI001, ASCKKWOH}@ntu.edu.sg

Abstract. Feature selection is an important preprocessing step in pattern recognition. In this paper, we presented a new feature selection approach in two-class classification problems based on information theory, named minimum Distribution Similarity with Removed Redundancy (mDSRR). Different from the previous methods which use mutual information and greedy iteration with a loss function to rank the features, we rank features according to their distribution similarities in two classes measured by relative entropy, and then remove the high redundant features from the sorted feature subsets. Experimental results on datasets in varieties of fields with different classifiers highlight the value of mDSRR on selecting feature subsets, especially so for choosing small size feature subset. mDSRR is also proved to outperform other state-of-the-art methods in most cases. Besides, we observed that the mutual information may not be a good practice to select the initial feature in the methods with subsequent iterations.

Keywords: Feature selection · Feature ranking · Information theory · Redundancy

1 Introduction

In many pattern recognition applications, the original dataset can be in a large feature size and may contain irrelevant and redundant features, which would be detrimental to the training efficiency and model performance [1, 2]. In order to reduce the undesirable effect of the curse of dimensionality and to simplify the model for parsimony [3], an intuitive way is to determine a feature subset, and this process is known as feature selection, or variable selection.

The ideal situation for feature selection is to select the optimal feature subset that maximize the prediction accuracy, however, this is impractical due to the intractable computation caused by the exhausted searching over the whole feature space, especially when the prior knowledge is limited and the dependency among features remains unknown. Therefore, varieties of suboptimal feature selection algorithms have been proposed, and they are mainly divided into three categories according to the evaluation metric: Wrapper, Embedded, and Filter methods [3].

Both Wrapper and Embedded methods are dependent on the classifiers. Wrapper methods score the feature subsets via the error rate on a given classifier, and certain searching strategies are employed to generate the next feature subset to avoid NP-hard problem [4, 5]. Such kinds of methods take the interactions among features into consideration and always can find the best subset for a particular learning algorithm, however, they are computation complex and prone to over-fitting, additionally, the subset needs to be reselected when changing the classifiers [6, 7]. In the meantime, embedded methods select the features during the model construction processes and thus they are more efficient than Wrappers [3], but the main limitation is that they rely heavily on the hypotheses the classifier makes [6, 8].

Unlike Wrapper and Embedded methods, Filter methods are independent of the classifiers, therefore they can better expose the relationships among features. Besides, Filter methods are simpler, faster and more scalable than Wrapper and Embedded methods, as they rank the features according to the proximity measures, such as the mutual information (MI) [3], correlation [10], chi-square [11] and relief-based algorithms [12]. The determination of the best feature subset of Filter methods is to select a cut-off point on their ranked features via the cross validation. But the drawback of Filter methods is that they cannot investigate the interaction between the features and classifiers [3, 5].

In Filter methods, information theory-based measure which exploiting not only the relationships between features and labels, but also the dependencies among features, plays a dominant role [9, 13]. Battiti [9] proposed Mutual Information Feature Selection (MIFS) method, which finds feature subset via greedy selection according to the MI between feature subsets and labels. After that, varieties of methods are presented to improve MIFS. Kwak and Choi developed MIFD-U method by considering more about the MI between features and labels [14], Peng *et al.* proposed minimal-redundancy-maximal-relevance (mRMR) framework by providing the theoretical analysis and combining with wrappers [15], Estévez *et al.* proposed Normalised MIFS (NMIFS) which replaces the MI with normalized MI [16], and Hoque *et al.* developed MIFS-ND by considering both MI of feature-feature and feature-label [17].

Consequently, a lot of information theory-based measures have been designed and adopted in Filter methods. Joint MI (JMI) [18], Interaction Capping (ICAP) [19], Interaction Gain Feature Selection (IGFS) [20] and Joint MI Maximisation (JMIM) [7] were raised by taking the joint MI into consideration, from which ICAP and IGFS depend on the feature interaction; Conditional MI Maximization (CMIM) criterion [21], Conditional Infomax Feature Extraction (CIFE) [22] and Conditional MIFS (CMIFS) [23] were proposed by adopting conditional MI; Double Input Symmetrical Relevance (DISR) method was developed by using symmetrical relevance as the objective function [24].

Intuitively, the common procedure for the above information theory related Filter methods is, selecting the initial feature with maximum MI, then increasing the feature subset size on previously defined features according to an objective function. Therefore, the selection of the initial feature will influence the determination of the final feature subset. As a result, a good initial feature may lead to a smaller feature subset size as well as a good classification performance. However, the maximum MI between the features and labels may not be a good criteria to determine the initial feature. We

argue that the initial feature found in this way may have a lower classification ability than many other features, which will further lead to relatively large feature subset size or a poor performance that actually can be avoided.

In this paper, we proposed a new approach to select feature subset based on information theory in two-class classification problems, named minimum Distribution Similarity with Removed Redundancy (mDSRR). Different from previous methods which use greedy search approach in adding the features into the feature subset, we rank the features according to their distribution similarity and then remove the redundancy from the ranked feature subset. Furthermore, we compared mDSRR with other state-of-the-art methods on 11 public datasets with different classifiers, results show that mDSRR is superior to other methods, which can achieve high performance with only a few features. Additionally, by comparing the classification performance and the initial feature defined by mDSRR and other methods, we demonstrated that using MI to determine the initial feature may not be a good practice.

This paper is organized as follows: Sect. 2 provides the background of information theory, theoretical analysis and the implementation of the proposed method, Sect. 3 demonstrates the results of the experiments and discuss the results, and Sect. 4 concludes this work.

2 Methods

2.1 Information Theory

This section briefly introduces the related concepts of information theory that will be used in this work. Suppose random variables X and Y represent feature vectors, and random variable C denotes the class label.

The entropy is a measure of the amount of uncertainty before a value of random variable is known, for a discrete random variable which takes value x from the alphabet χ , i.e. $x \in \chi$, with probability $p_X(x)$, the entropy is defined as

$$H(X) = - \sum_{x \in \chi} p_X(x) \log(p_X(x)) \quad (1)$$

The entropy is positive and bounded, i.e. $0 \leq H(X) \leq \log(|X|)$. Similarly, when taking two discrete random variables X and Y and their joint probability $p(x, y)$ into consideration, the joint entropy can be represented as

$$H(X, Y) = - \sum_{x \in \chi} \sum_{y \in \mathcal{Y}} p(x, y) \log(p(x, y)) \quad (2)$$

The conditional entropy of X given that C is a measure of the average additional information in X when C is known, which is defined as

$$H(X|C) = - \sum_{c \in \mathcal{C}} \sum_{x \in \chi} p(x, c) \log(p(x|c)) \quad (3)$$

where $p(x|c)$ is the conditional probability for x given that c , and according to the chain rule,

$$H(X, C) = H(X|C) + H(C) \quad (4)$$

Mutual information of X and Y is the average amount of information that we get about X from observing Y , or in other word, the reduction in the uncertainty of X due to the knowledge of Y , it is represented as

$$I(X; Y) = H(X) - H(X|Y) = H(X) + H(Y) - H(X, Y) \quad (5)$$

The mutual information is symmetrical and non-negative, it equals to 0 only when the variables are statistically independent. Similarly, the conditional mutual information of X and Y given that C is represented as

$$I(X; Y|C) = H(X|C) - H(X|Y, C) \quad (6)$$

Kullback-Leibler divergence, also known as relative entropy, is another important concept that will be used in this work. It measures the difference between two probability mass vectors, if we denote the two vectors as \mathbf{p} and \mathbf{q} , then the relative entropy between \mathbf{p} and \mathbf{q} is defined as

$$D(\mathbf{p}||\mathbf{q}) = \sum_{x \in \mathcal{X}} p(x) \log \left(\frac{p(x)}{q(x)} \right) \quad (7)$$

Obviously, it is asymmetric between \mathbf{p} and \mathbf{q} .

2.2 Minimum Distribution Similarity Feature Ranking

We try to rank the features according to their distribution similarities between classes to separate the objects, where the distribution similarity can be measured by the relative entropy.

Recalling the mathematic representation of relative entropy, Eq. (7) in Sect. 2.1, $D(\mathbf{p}||\mathbf{q})$ represents the “distance” between the probability mass vectors \mathbf{p} and \mathbf{q} , and it also can be regarded as the measurement of information loss of \mathbf{q} from \mathbf{p} . Here, we regard x as the event representing the instances in one type of feature whose values located in a certain range, or a small bin, X as the event set, and \mathbf{p} and \mathbf{q} as the distributions of x for two classes separately. In this work, we use small bins rather than using the exact value of feature, because when $q(x) = 0, p(x) \neq 0, \log(\frac{p(x)}{q(x)}) \rightarrow \infty$; when $p(x) = 0, q(x) \neq 0, p(x) \log(\frac{p(x)}{q(x)}) = 0$. Obviously, if there is no overlapped feature values in two classes, $p(x) \log(\frac{p(x)}{q(x)})$ would either equal to ∞ or 0, under this circumstance, $D(\mathbf{p}||\mathbf{q})$ would equal to ∞ or 0. It is worth noting that if a large number of features do not have overlapped values in different classes, all these features would be assigned value of infinity or zero and we cannot know which one is more important.

Although introducing small bins can solve the above problem, if the bin number is selected too small, the difference for two classes cannot be captured accurately; if it is too large, the relative entropy value would tend to be the same as when taking exact value stated above. Therefore, several choices of bin number determined according to the corresponding total instance amount are evaluated to select the proper solution here. Furthermore, since $D(p\|q) \neq D(q\|p)$, we use $D = D(p\|q) + D(q\|p)$ as the measure to sort the features.

Another problem is how to deal with the circumstance when $D \rightarrow \infty$ in practical. In this work, we use $q(x) = \frac{1}{\text{instance number of class II}}$ in $D(p\|q)$ to replace $q(x) = 0$ to avoid D becomes infinity. This practice, although may not so accurate, can reflect the trend of D . On one side of the spectrum, when the number of total instances in a certain bin is large, the measure of distribution similarity when there is only one sample of class II in this bin is almost the same as that when there is no sample of class II, both of the two circumstance means an extremely low similarity between the distribution for two classes. At another extreme, when the instance number in a certain bin is small and there is only samples of class I in this bin, due to its small sample number, $p(x)$ tends to be 0, $p(x) \log\left(\frac{p(x)}{q(x)}\right)$ would be an extremely small value and would not contain too much information. Therefore, replacing $q(x) = 0$ with $q(x) = \frac{1}{\text{instance number of class II}}$ in the practical calculation is rational. This practice can reflect the real distribution similarity trend, especially under the truth that this circumstance would not happen too much with the application of bin. The above algorithm has been briefly described in a previous work [25].

2.3 Minimum Distribution Similarity with Removed Redundancy (mDSRR)

In Sect. 2.2, we already got a feature list sorted according to their potential importance to classification, where the irrelevant features will be ranked in the back. However, the redundancy may still exist. The combination of the features redundant to each other may not contribute to higher performance but lead to overfitting. Therefore, removing redundant features is necessary to improve model's performance and efficiency. Different from previous works which use MI or conditional MI between features and label, in this work, we consider the conditional MI between features under the condition of label as the criteria to remove redundancy. In particular, we adopt conditional MI rather than the MI because according to, features have high mutual information may have different information within the class, hence taking labels into consideration is essential.

The complexity to calculate the conditional MI between two features within a feature set with feature size n is proportional to $(n-1) + (n-2) + \dots + 1 = \frac{n^2-n}{2}$, as the increase of n , the calculation complexity will grow as n^2 . In case of the feature number is large, it would be impractical to find all conditional MI as the time consumption would be large. Hence, we only calculate the conditional MI for the first m ranked features.

From Eq. (4) and (6), we can obtain

$$\begin{aligned} I(X; Y|C) &= H(X|C) - H(X|Y, C) = H(X|C) - H(X, Y|C) + H(Y|C) \\ &= H(X|C) + H(Y|C) + H(C) - H(X, Y, C) \end{aligned} \quad (8)$$

to calculate the conditional MI between two features under the condition of labels. The relationship among the items in Eq. (8) is illustrated in Fig. 1. From the Venn diagram, the redundancy between feature X and Y under the condition of class C can be measured as the percentage of $I(X; Y|C)$ taken in $H(X, Y|C)$, which can be represented as the ratio r between $I(X; Y|C)$ and $H(X|C) + H(Y|C) - I(X; Y|C)$, that is

$$r = \frac{I(X; Y|C)}{H(X|C) + H(Y|C) - I(X; Y|C)} \quad (9)$$

where $0 \leq r \leq 1$. In Eq. (9), $r = 0$ means X and Y are independent under the condition of C; the larger the r is, the larger the redundancy between X and Y; when $r = 1$, X and Y are totally dependent, or totally redundant under the condition of the label. And once r exceeds the predefined threshold, the feature which is ranked in later position between two redundant features will be removed from the feature subset.

Algorithm 1. mDSRR method

```

1 Input: class I data feature set  $P$ , class II data feature set  $Q$ , bin number  $n$ .
2 Initialize: initial feature set  $F$ .
3 Begin
4   for  $i = 1, \dots, |P|$ :
5      $D_i = D(p_i||q_i, n) + D(q_i||p_i, n)$ 
6   end
7   sort  $D_i$  from largest value to smallest value, get a new feature set  $F_{new}$ .
8   take the first  $m$  items in  $F_{new}$ ,  $\rightarrow F_m$ .
9   for  $j = 1, \dots, m$ :
10    calculate  $H(F_m(j)|C)$ ,  $H(F_m(j+1)|C)$ , and  $I(F_m(j); F_m(j+1)|C)$ 
11     $r_{j,j+1} = \frac{I(F_m(j); F_m(j+1)|C)}{H(F_m(j)|C) + H(F_m(j+1)|C) - I(F_m(j); F_m(j+1)|C)}$ 
12    if  $r_{j,j+1} > r_{th}$ :
13      set  $F_m \leftarrow F_m \setminus \{f_m(j+1)\}$ 
14    end
15  end
16 End
17 Output:  $F_m$ .

```

* $|P| = |Q|$ is the number of features.

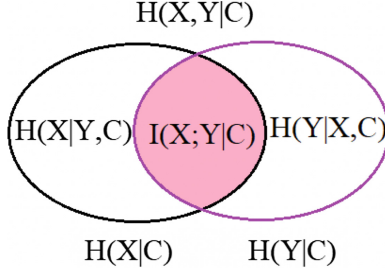


Fig. 1. Venn diagram of the relationship among items in Eq. (8).

Combining the feature ranking algorithm in Sect. 2.2 and the remove redundancy step described above, our feature selection method is finally developed and named as mDSRR (minimum Distribution Similarity with Removed Redundancy). The algorithm is summarized in Algorithm 1.

3 Results

To highlight the effectiveness of mDSRR in feature selection, five state-of-the-art methods, including mRMR [15], DISR [24], ICAP [19], CIFE [22], and CMIM [21] are used for comparison. These five approaches are chosen because that, (i) they are all Filter methods based on information theory; (ii) they cover information theory measures like MI, joint MI and conditional MI; (iii) they are classic and popular feature selection methods which have been applied widely in diversity of areas. All methods are evaluated on 11 public datasets and 4 kinds of classifiers, including Supporting Vector Machine (SVM), Decision Tree (DT), Random Forest (RF) and Naïve Bayes (NB). The average classification accuracy (Acc) for the 10-fold cross validation, which can reflect the general performance while avoiding the bias, is recorded and works as the criteria to evaluate different methods.

3.1 Datasets

Eleven public datasets from UCI Repository [26] are used for comparison, they were all in two classes with multivariate and integer or real attributes, and covering a diversity of areas, such as life, economic, chemistry and biology, medical, computer, artificial and physical. These datasets vary in instance numbers and feature numbers, hence can be used for fair comparisons, the related information for these datasets are briefly listed in Table 1.

Table 1. Datasets used in this work.

Dataset name	Number of instances	Number of features
Arcene	200	10000
Audit	776	17
Biodegradation	1055	41
Breast Cancer Wisconsin	683 (origin 699)	10
Breast Cancer Coimbra	116	9
Diabetic Retinopathy Debrecen	1151	19
Madelon	2600	500
Musk	7074	166
Parkinson	756	751
Sonar	208	60
Spambase	4601	57

3.2 Parameter Determination in mDSRR

In mDSRR, the parameter bin number needs to be assured, to achieve it, we did experiments on 11 datasets to see the impacts of the selection of bin number. Following the rules what we described in Sect. 2.2, the bin number cannot be selected too large and too small, we consider the circumstance when bin number equals to $\frac{1}{10}, \frac{1}{15}, \frac{1}{20}, \frac{1}{30}$ and $\frac{1}{50}$ of the total number of instance for dataset with a relatively large instance amount, i.e. > 600 , and $\frac{1}{2}, \frac{1}{5}, \frac{1}{10}, \frac{1}{15}$ and $\frac{1}{20}$ of the total number of instance for dataset with a relatively small instance amount, i.e. < 300 . The comparison of the Acc achieved by feature subsets with different choices of bin number for 11 datasets are plotted in Supplementary Figure S1 and S2. The best choice of bin number is selected as the one that achieves the best Acc most times in four classifiers, if two or more

Table 2. The best choice of bin number for 11 datasets.

Datasets	Number of instances	Best bin number in portion*
Breast Cancer Coimbra	116	1/5
Arcene	200	1/15
Sonar	208	—**
Breast Cancer Wisconsin	683	—
Parkinson	756	1/20
Audit	776	1/20 or 1/30
Biodegradation	1055	1/50
Diabetic Retinopathy Debrecen	1151	1/50
Madelon	2600	1/50
Spambase	4601	1/50
Musk	7074	1/50

*Best bin number in portion is the portion of bin number taken up in total instance number.

**Means the choice of best bin number percentage is hard to define according to the existing results.

choices of bin number realize the best Acc equal times, then we consider their performance under the same feature subset size. The times for different choices of bin number which achieving the best Acc are counted in Supplementary Table S1, and the summarize of the best choices of bin number for these datasets are listed in Table 2.

From the experiment results, the portion that the bin number taking up in the total instances decreased as the total number of instances increases. For datasets with instance number at around 100, we use bin number roughly as 1/5 of the total instance number; for total instance number around 200–500, we use its 1/15 as bin number; for total instance number around 500–1000, we use 1/20; for total instance number larger than 1000, we use 1/50.

Actually, with the histogram idea, there is no best choice of bin number without a strong assumption about the shape of the distribution, and the parameter we chose is a suboptimal one, but in later sections, we will show the good performance of mDSRR with such choices of bin number.

The other parameter in mDSRR is the redundancy remove threshold r_{th} . Although sometimes the redundancy between two features may be large, they may still contain useful information and should not be deleted, hence we consider the circumstances when $r_{th} = 0.9, 0.99$ and 0.9999 . The feature subset sizes after removing redundancy with different selection of r_{th} are recorded in Supplementary Table S2 and their comparisons are plotted in Supplementary Figure S3. 0.9999 is chosen as the final value of r_{th} which achieves the best performance most times, although in some cases no feature is removed with such a high threshold.

3.3 Performance Comparison on Datasets with Large Feature Size

To compare the performance of different feature selection methods on datasets with a relatively large feature set size, we plotted the average Acc of 10-fold cross validation for feature subsets with different sizes in Fig. 2. For datasets whose feature size exceeding 100, e.g. Arcene, Madelon, Musk and Parkinson, the results of feature subsets with size from 1 to 50 are shown, and for the others (Biodegradation, Spambase and Sonar), the results of feature subsets with all possible sizes are shown. The bin numbers are chosen following the rules we concluded in Sect. 3.2.

The application of mDSRR method on dataset Arcene, Parkinson and Sonar has obvious advantages over other methods, the feature subset determined by mDSRR can achieve the highest value with only a small feature subset size. For example, in Arcene dataset, the feature subset with size 10 selected by mDSRR realize 85% Acc with SVM classifier, while other methods never reach this value no matter how many features are added; although mDSRR does not realize the best Acc in this dataset with NB classifier, the feature subset it determined with only 2 features achieve Acc as high as 71%, which is almost the same as the best Acc achieved by ICAP (71.5%), with feature subset size 32. Furthermore, with the same feature subset size which are less than 25, 40 and 10 in dataset Arcene, Parkinson and Sonar, respectively, the performance of mDSRR are much better than other methods in most cases.

Although the general performances of mDSRR on the remaining 4 datasets is not as outstanding as the above datasets, the advantages of mDSRR still can be found. For Biodegradation dataset, the best Acc are achieved by mRMR with SVM and DT

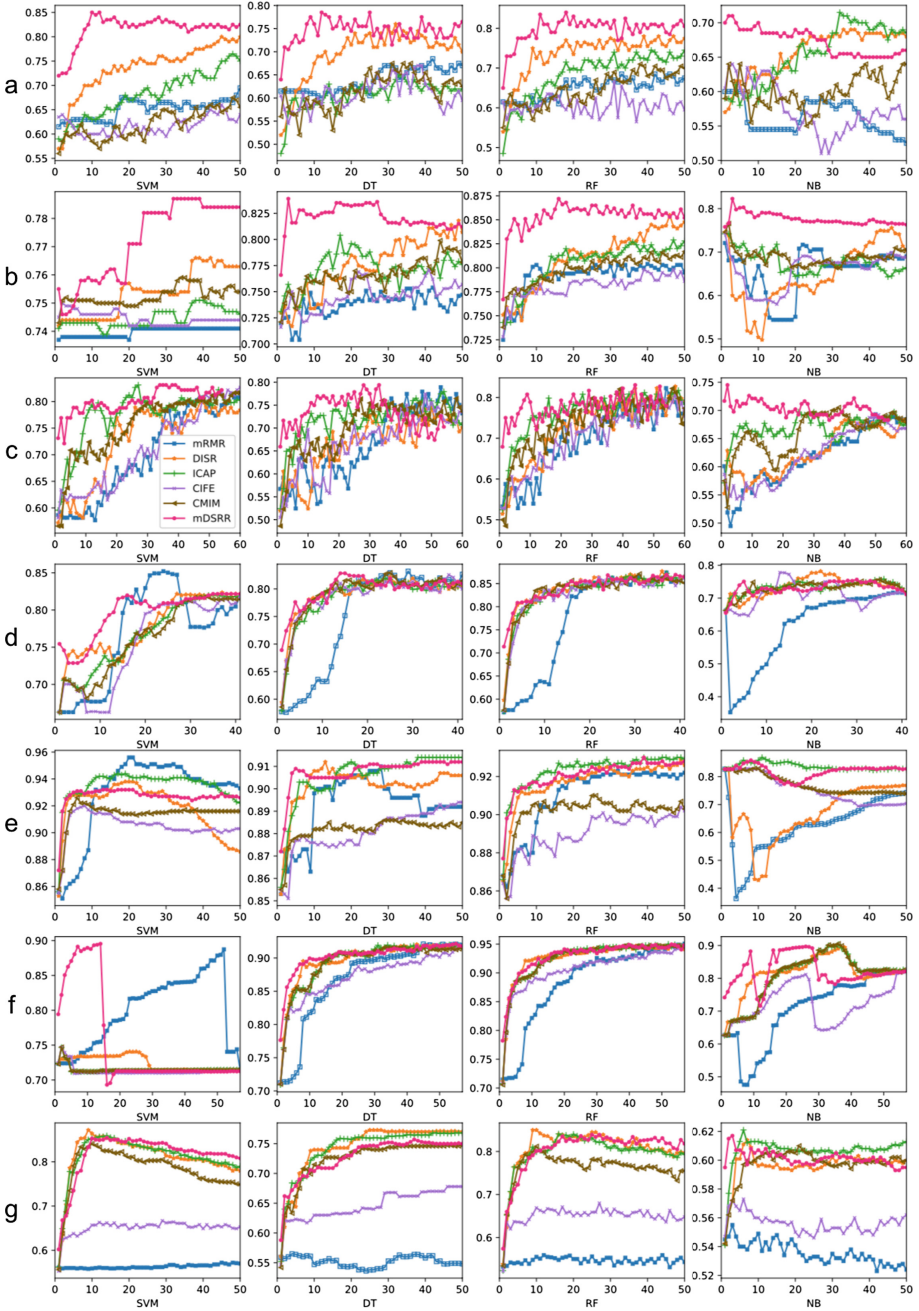


Fig. 2. The average Acc of different feature subset sizes for different feature selection methods with classifier SVM, DT, RF and NB on dataset (a) Arcene, (b) Parkinson, (c) Sonar, (d) Biodegradation, (e) Musk, (f) Spambase, and (g) Madelon.

classifier, but when the feature number is smaller than 16 and 18 in two classifiers separately, the Acc value of mDSRR are higher than that of mRMR with up to 9.7% and 18.7% within the same feature subset sizes. Additionally, mDSRR reaches 75.1% Acc with only 5 features in NB classifier, while DISR, the one realizes the best Acc, reaches the same value with 17 features. For Musk dataset, mDSRR achieves Acc of 93% and 90.9% in SVM and DT separately with only 5 features, while mRMR and ICAP, the methods that achieve the best Acc in two classifiers, reach the same value with 10 and 17 features separately. For Spambase dataset, mDSRR achieves the best Acc of 89.5% with only 14 features in SVM classifier, and when the feature subset sizes are smaller than 18 and 29 in DT and NB classifiers, mDSRR leads to much higher performances than other methods. And for Madelon dataset, the feature subset determined by mDSRR realizes the second-highest Acc value with only 3 features in NB classifier, which far exceeding the performance with the same feature subset size defined by other methods.

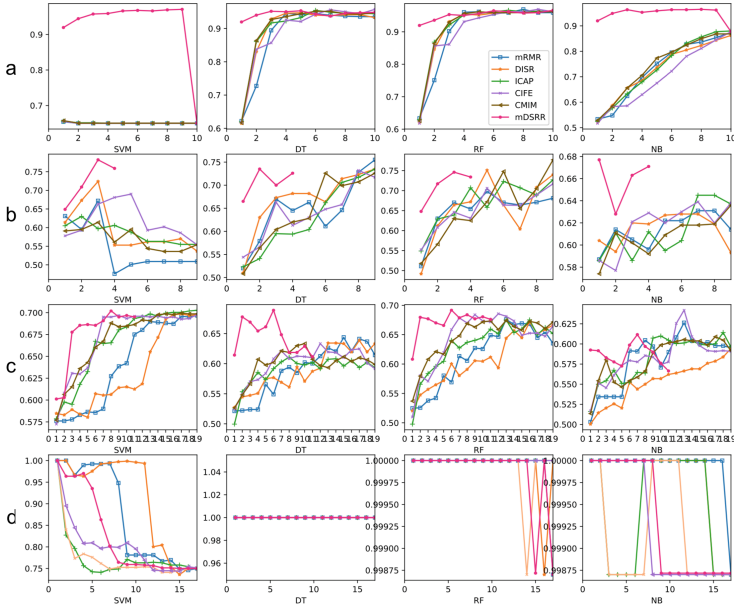


Fig. 3. The average Acc of different feature subset sizes for different feature selection methods with classifier SVM, DT, RF and NB on dataset (a) Breast Cancer Wisconsin, (b) Breast Cancer Coimbra, (c) Diabetic Retinopathy Debrecen, and (d) Audit.

3.4 Performance Comparison on Datasets with Small Feature Size

For datasets with a relatively small feature number, the impact of a single feature can be significant. We plotted the average Acc achieved by 10-fold cross validation obtained by the feature subsets with different sizes selected by different methods for four datasets whose feature set sizes are relatively small, as shown in Fig. 3.

The overall performance of mDSRR is much better than the other methods on all datasets except Audit, where mDSRR achieves the highest Acc with only several features in most cases. For instances, in Breast Cancer Wisconsin dataset, the Acc for feature subsets with size from 1 to 9 selected by mDSRR remain at a high Acc value, i.e. no less than 92%, while the feature subsets determined by other methods only reach this value when the subset size is larger than 4 for DT and RF classifiers, but never exceed 90% for SVM and NB classifiers. In Breast Cancer Coimbra dataset, the remove redundancy step in mDSRR removes 2 to 5 features in 10-fold split datasets, hence we only kept the first 4 features for plotting. The feature subset selected by mDSRR with size 3 in SVM and size 1 in NB classifier achieve the best values, while the feature subsets determined by other methods never reach the same values no matter how many features are used. Similarly, in Diabetic Retinopathy Debrecen dataset, mDSRR removes one feature and it achieves the best Acc with only 8 features in SVM classifier, and with 6 features in DT and RF classifiers, which is superior to other methods with better performance but smaller feature subset size. In addition, for the above three datasets, when feature subset sizes are same and less than a certain value, mDSRR keep leading to higher performance than other methods. As to the Audit dataset, the first feature selected by all methods remains the same which realize 100% Acc, although the performance for later feature subsets may vary, all feature selection methods can be regarded as performing equally.

3.5 Evaluation of the Ability of MI to Find the Initial Feature

Most Filter feature selection methods enlarge their feature subsets on the initial feature according to a loss function, hence the selection of the initial feature is extremely important. The common practice to select the initial feature is to find the one that maximizes the MI between the feature and classes, i.e. $\max\{I(X_i; C)\}$, as this solution is proved to be near the optimal to minimize the Bayes error [22, 27]. All 5 methods used for comparison in this work employ MI to identify their initial feature. However, determining the initial feature according to MI may not be a good practice in practical, the initial feature chosen in this way may have a poor classification performance or its performance is worse than that of other features, under this circumstance, the feature subset chosen followed by this feature would perform poor either or require more features to reach the same result as those determined with a good initial feature.

A good example to show the weakness of MI as the criteria to select the initial feature in those subsequent iteration methods is Breast Cancer Wisconsin dataset. In this dataset, mDSRR determines the 4th feature and the 1st feature in the original feature set as the first and the last feature to be added to the feature subset separately, while the other 5 methods select the 1st feature in original feature set as the initial feature. However, combining the results in Sect. 3.4, the 1st feature in original feature set leads to poor performances with SVM and NB classifiers, whose Acc are around 65% and 53% separately, and when this feature is added to the feature subset found by mDSRR, the high performances of other features, i.e. over 90% Acc, decrease significantly. Particularly, when the 1st feature exists, the accuracies of SVM and NB classifiers can never exceed 65% and 90% separately. Hence, when using the initial feature identified by MI in Breast Cancer Wisconsin dataset, the performance with SVM and NB

classifier can hardly be improved. The finding here proved that it is important to properly select the initial feature for methods which based on the subsequent iteration, and the criteria of maximum MI may not be a good choice to determine the initial feature.

4 Conclusion

In this paper, we proposed a new feature selection method based on information theory: minimum Distribution Similarity with Removed Redundancy (mDSRR). mDSRR employs the concept of relative entropy, combined with the histogram idea to rank the features, and the redundancy is removed according to the conditional MI between two features under the condition of classes. The idea of mDSRR is different from previous information theory related Filter feature selection method, which usually follow the practice of determining an initial feature and then enlarging the feature subset on the initial feature according to a loss function.

The comparison results between mDSRR and five state-of-the-art methods on 11 public datasets with four kinds of classifiers in this work show that mDSRR is a valuable method to select effective feature subset. mDSRR leads to better performance than other methods under the same sizes of the feature subsets especially when the sizes are small. Besides, by taking Breast Cancer Wisconsin dataset as an example, we also demonstrate that MI may not be a good practice to determine the initial feature in those subsequent iteration methods.

However, one limitation of mDSRR is that it can only be utilized to two-class classification problems, while are not suitable for multi-class classification problems. Because the relative entropy only can calculate the “distance” between two distributions. For multi-class classification problems, if we measure the distribution similarity of any two classes and then integrate the results, the workload would be heavy.

In a nutshell, mDSRR is a good method to select feature subset, especially to select small size feature subset due to its high efficiency in ranking the features according to their potential contribution to distinct the classes. The successful applications of mDSRR on a range of datasets in different fields with different classifiers highlight its value in feature selection.

Supplementary Data

Supplementary data are available at https://github.com/yuuuuzhang/feature-selection/blob/master/fs_supplementary.docx.

References

1. Yan, H., Hu, T.: Unsupervised dimensionality reduction for high-dimensional data classification. *Mach. Learn. Res.* **2**, 125–132 (2017)

2. Gu, S., Cheng, R., Jin, Y.: Feature selection for high-dimensional classification using a competitive swarm optimizer. *Soft. Comput.* **22**(3), 811–822 (2018)
3. Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. *J. Mach. Learn. Res.* **3**, 1157–1182 (2003)
4. Chandrashekar, G., Sahin, F.: A survey on feature selection methods. *Comput. Electr. Eng.* **40**(1), 16–28 (2014)
5. Wah, Y.B., Ibrahim, N., Hamid, H.A., Abdul-Rahman, S., Fong, S.: Feature selection methods: case of filter and wrapper approaches for maximising classification accuracy. *Pertanika J. Sci. Technol.* **26**(1), 329–340 (2018)
6. Jain, D., Singh, V.: Feature selection and classification systems for chronic disease prediction: a review. *Egypt. Inform. J.* **19**(3), 179–189 (2018)
7. Bennasar, M., Hicks, Y., Setchi, R.: Feature selection using joint mutual information maximisation. *Expert Syst. Appl.* **42**(22), 8520–8532 (2015)
8. Hira, Z.M., Gillies, D.F.: A review of feature selection and feature extraction methods applied on microarray data. *Adv. Bioinform.* **2015**, 1–13 (2015)
9. Battiti, R.: Using mutual information for selecting features in supervised neural net learning. *IEEE Trans. Neural Netw.* **5**(4), 537 (1994)
10. Liu, H., Li, J., Wong, L.: A comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns. *Genome Inform.* **13**, 51–60 (2002)
11. Jin, X., Xu, A., Bie, R., Guo, P.: Machine learning techniques and chi-square feature selection for cancer classification using SAGE gene expression profiles. In: Li, J., Yang, Q., Tan, A.-H. (eds.) *BioDM 2006. LNCS*, vol. 3916, pp. 106–115. Springer, Heidelberg (2006). https://doi.org/10.1007/11691730_11
12. Urbanowicz, R.J., Meeker, M., La Cava, W., Olson, R.S., Moore, J.H.: Relief-based feature selection: introduction and review. *J. Biomed. Inform.* **85**, 189–203 (2018)
13. Torkkola, K.: Feature extraction by non-parametric mutual information maximization. *J. Mach. Learn. Res.* **3**, 1415–1438 (2003)
14. Kwak, N., Choi, C.-H.: Input feature selection for classification problems. *IEEE Trans. Neural Netw.* **13**(1), 143–159 (2002)
15. Peng, H., Long, F., Ding, C.D.: Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Patter Anal. Mach. Intell.* **1**(8), 1226–1238 (2005)
16. Estévez, P.A., Tesmer, P.A., Perez, C.A., Zurada, J.M.: Normalized mutual information feature selection. *IEEE Trans. Neural Netw.* **20**(2), 189–201 (2009)
17. Hoque, N., Bhattacharyya, D.K., Kalita, J.K.: MIFS-ND: a mutual information-based feature selection method. *Expert Syst. Appl.* **41**(14), 6371–6385 (2014)
18. Yang, H., Moody, J.: Feature selection based on joint mutual information. In: *Proceedings of International ICSC Symposium on Advances in Intelligent Data Analysis* (1999)
19. Jakulin, A.: Machine learning based on attribute interactions. *Univerza v Ljubljani* (2006)
20. Akadi, A.E., Ouardighi, A.E., Aboutajdine, D.: A powerful feature selection approach based on mutual information. *Int. J. Comput. Sci. Netw. Secur.* **8**(4), 116 (2008)
21. Fleuret, F.: Fast binary feature selection with conditional mutual information. *J. Mach. Learn. Res.* **5**, 1531–1555 (2004)
22. Lin, D., Tang, X.: Conditional infomax learning: an integrated framework for feature extraction and fusion. In: Leonardi, A., Bischof, H., Pinz, A. (eds.) *ECCV 2006. LNCS*, vol. 3951, pp. 68–82. Springer, Heidelberg (2006). https://doi.org/10.1007/11744023_6
23. Cheng, G., Qin, Z., Feng, C., Wang, Y., Li, F.: Conditional mutual information-based feature selection analyzing for synergy and redundancy. *ETRI J.* **33**(2), 210–218 (2011)

24. Meyer, P.E., Bontempi, G.: On the use of variable complementarity for feature selection in cancer classification. In: Rothlauf, F., et al. (eds.) *EvoWorkshops 2006*. LNCS, vol. 3907, pp. 91–102. Springer, Heidelberg (2006). https://doi.org/10.1007/11732242_9
25. Zhang, Y., Jia, C., Fullwood, M.J., Kwoh, C.K.: DeepCPP: a deep neural network based on nucleotide bias and minimum distribution similarity feature selection for RNA coding potential prediction. *Brief. Bioinform.* (2020). <https://doi.org/10.1093/bib/bbaa039>
26. Dua, D., Graff, C.: UCI Machine Learning Repository (2019). <http://archive.ics.uci.edu/ml>
27. Vasconcelos, N.: Feature selection by maximum marginal diversity. In: *Advances in Neural Information Processing Systems*, pp. 1375–1382 (2003)