ISSN: 0976-2876 (Print)
ISSN: 2250-0138(Online)

A STUDY OF A NEW FEATURE SELECTION ALGORITHM BY MAXIMIZING INDEPENDENT CLASSIFICATION INFORMATION

[1]Mr.V Mohan, [2]Ms. KoustubhaMadhavi B, [3]Dr.S.B.Kishor
[1,2] Department of Computer Science and Engineering, NallaMalla Reddy Engineering College, Hyderabad, Telangana.
[3]Department of Computer Studies and Research, S.P College, Chandrapur, Maharsashtra.

Abstract- Most real world datasets contain a certain degree of redundancy in the form of identical object instances, non relevant features and features that are dependent on one another. In a data mining context, this redundancy can lead to the extraction of spurious rules and can make learning very expensive in a classifier system. Feature selection aims at filtering out the irrelevant features and can be viewed as a pre-processing step in knowledge extraction or classifier training. Feature selection algorithms have been applied to datasets of a wide variety of fields such as image recognition, bioinformatics, text classification, text clustering etc. Most of the feature selection algorithms work with labeled datasets and also require some kind of subjective inputs from the user. On the other hand, unsupervised feature selection algorithms work with unlabeled datasets. Feature selection approaches based on mutual information can be roughly categorized into two groups. The first group minimizes the redundancy of features between each other. The second group maximizes the new classification information of features providing for the selected subset. A critical issue is that large new information does not signify little redundancy, and vice versa. Features with large new information but with high redundancy may be selected by the second group, and features with low redundancy but with little relevance with classes may be highly scored by the first group. Existing approaches fail to balance the importance of both terms. In this paper, we study and present a new information term denoted as Independent Classification Information. It assembles the newly provided information and the preserved information negatively correlated with the redundant information. This strategy helps find the predictive features providing large new information and little redundancy.

Keywords -Feature Selection , Independent Classification Information , Feature Redundancy

I. Introduction

A feature selection algorithm can be used to classify the feature subsets which are identified and removed as much of the irrelevant and redundant information as possible, along with an evaluation measure. The best subset contains the least number of dimensions that most contributed to accuracy. The feature selection is important to speed up training and to improve generalization performance[1]. In this active field of research, numerous classic feature selection algorithms have been widely-used, such as wrappers, filters and embedded methods[2]. Filter methods use a measure to capture the usefulness of the feature subsets from the high-dimension data sets, for example, using the common measures which based on the mutual information, it can allow the feature selection algorithms to operate faster and more effectively. The traditional feature selection algorithms use Shannon's mutual information (MI) as a measure of relevance among features. But the MI method has the disadvantages of redundancy. In 1994, Battiti [11] proposed mutual information feature selection (MIFS) which selected the feature that maximizes the information about the class, corrected by subtracting a quantity proportional to the average MI with the previously selected features. Kwak and Chan [4] analyzed the limitations of MIFS and proposed a greedy selection method called MIFS-U, which in general, makes a better estimation of the MI between input attributes and output classes than MIFS. In view of the above analysis, a new information term,

Independent Classification Information (ICI), is studiedinthis paper. It unifies redundancy information and new classificationinformation in one term. Thus, the importance ofthese two kinds of information is synthetically consideredby ICI. Two kinds of conditional mutual information areemployed by ICI to evaluate the contributions of candidateand selected features for classification. One kind of informationis newly provided by a candidate feature, whichdenotes the particular

[1]Corresponding Author

contribution of this feature differentfrom that of the selected features. Another kind of informationis preserved by the selected features if a candidatefeature is selected. This information represents the particularcontributions of these features that is different fromthe candidate feature, and exhibits a negative correlationwith the feature redundancy for classification. Therefore,ICI focuses on the differences between features in their classificationabilities. This strategy helps find highly discriminativeas well as lowly redundant features. ICI is alsoproved as a loose upper bound of the global classificationinformation of feature subset. Thus, the new method isexpected to obtain a high global classification performance.

The paper is organized as follows. In Section II the fundamentals of Mutual Information among features is discussed. The concept of independent classification information is introduced in Section III. Section IV demonstrates the classification comparison is done on various datasets.

II.Background Study

A.Mutual Information

Feature selection is a critical technology to reduce dimensionality. It helps prevent the curse of dimensionality and extract a good representation of the original variable model. Selection methods are typically divided into supervised, semi-supervised, and unsupervised [29]. Supervised methods such as Laplacian Score [30], Inf-FS [31], ReliefF [32] employ class labels to measure the discriminative abilities of features.Mutual Information [6] is used to quantitatively analyze the mutual dependence between any two features or between a feature and a class variable. The mutual information of two continuous random variables X and Y is an effective criterion to measure variable correlation [1]. The mutual information between two variables y and x is defined as :

$$I(y;\, x)=H(y)-H(y|x)\,\mathbf{(1)}$$

Where $H(y)$ and $H(y|x)$ represent the entropy and conditional entropy of the involved variables. It describes the decreased uncertainty for one variable when another variable is given, that is, their shared information [2]. Mutual information is widely utilized to evaluate the discriminative performance of features [3], [4]. These methods aim to find the most relevant features [5], [6] to the target class [7]. This mechanism can be denoted as the maximization of Eq. (2), supposing features x_1 ,$x_{2\ldots\ldots}x_k$ are evaluated and y is the target class for recognition:

$$I(y;\, x, \ldots x_k) = H(y)-H(y|x_{1}, \ldots\ldots x_k) \ldots\ldots\ldots\ldots\ldots\mathbf{(2)}$$

The features maximizing Eq. (2) are recognized as most discriminative for y because of their maximal information for classification. Theoretically, it can be calculated as:

$$I(y\colon x_1, ..)=\sum_y \sum_{x_i} \sum_{x_k} P(\,y\colon x_1, .., x_k) log \frac{P(y\colon x_1,..,x_k)}{P(y)P(y\colon x_1,..,x_k)}\mathbf{(3)}$$

In all related work, including the mutual information-based methods, how to select informative features while reducing feature redundancy is an important issue to be addressed all along. Intuitively, mutual information can be directly applied to feature selection by maximizing the relevance of candidate feature x_k with classes, which is represented by the Max-Relevance criterion as follows:

$$J_{Max\_Rel}(x_k) = I(y; x_k)\mathbf{(4)}$$

Discriminative but redundant features are selected by Max-Relevance, and thus result in inferior performance to the expected outcome in the recognition task. Therefore, the issue of alleviating redundant information receives more attention [33]. Two representative methods, namely,MIFS [11] and mRMR [12], are proposed as follows, supposing the feature subset

$S = \{x_1, .., x_{k-1})$is selected

$$J_{MIFS}(x_k) = I(y; x_k) - \beta \sum_{x_j \in S} I(x_j; x_k)\mathbf{(5)}$$

$$J_{mRMR}(x_k) = I(y; x_k) - \frac{1}{|S|} \sum_{x_j \in S} I(x_j; x_k)\mathbf{(6)}$$

Feature redundancy is reduced by both methods, in which the mutual information of two features is directly considered as their redundancy and minimized.$I(x_j; x_k)$quantifies the amount of information that two features share, which may or may not be relevant to classification. Obviously, only the information shared by two features to recognize class $\boldsymbol{y}$ should be regarded as redundant for classification. This information is de facto the multi-information $I(y; x_j; x_k)$ in Eq. (6). $I(y; x_j; x_k)$ can also be computed as $I(y; x_j; x_k) = I(y; x_k) - I(y; x_k|x_j)$ [26]. This implies that information provided by x_k partially contributes to classification, because this information also involves the redundant information possessed by the selected feature x_j . Note that $I(y; x_j; x_k)$ may obtain both positive and negative values . It is positive if adding the condition feature x_j reduces the relevance of x_j with y, which can be interpreted as the class-relevant redundancy of two features. Conversely, a negative value is obtained if adding

x_j helps enhance this relevance. In this case, two features are complementary for recognition. Some methods, such as CIFE [14], MIFS-U [15], CMIFS [16], ICAP [17], mIMR [18], and IGFS [19], employ multi-information in their evaluation criteria to determine the redundancy of two features. The criteria of CIFE and ICAP are shown as follows:

$$J_{CIFE}(x_k) = I(y; x_k) - \sum_{x_j \in S} I\big(y; x_j : x_k\big) \textbf{(7)}$$

$$J_{ICAP}(x_k) = I(y; x_k) - \sum_{x_j \in S} \max\,[0, I\big(x_j; x_k\big)] \textbf{(8)}$$

Reducing redundancy can enhance the discriminative ability of a feature subset. A more direct way is to maximize the classification information newly provided for feature subset by candidate features. The joint mutual information between the subset and classes is expected to be increasedby this strategy. JMI [22], IF [23], DISR [24] and CMIM [25] can be included into this group. In contrast to redundancy reduction methods, which take the target *y* as a condition, the selected features are considered as conditions in these methods. JMI in Eq. (9) and CMIM in Eq. (10) illustrate this idea:

$$J_{JMI}(x_k) = \sum_{x_j \in S} I\big(x_k, x_j, y\big) \propto \sum_{x_j \in S} I\big(y; x_j : x_k\big) \textbf{(9)}$$

$$J_{CMIM}(x_k) = min_{x_j \in S}\big[I\big(y; x_j : x_k\big)\big] \textbf{(10)}$$

$I\big(y; x_j : x_k\big)$ quantifies the amount of the classification information that x_k provides when x_j has been selected [34]. This information cannot be provided by *S*. Compared with$I(y; x_k), I\big(y; x_j|x_k\big)$ does not involve the redundant information of pair wise features for classification. Some methods, which aim to reduce redundancy, can be transformed into the methods that select features with large new classification information according to Eq. (9) [26]. When examining a candidate feature x_k , increasing $I\big(y; x_k|x_j\big)$ is equivalent to decreasing $I\big(y; x_j|x_k\big)$. However, $I\big(y; x_k|x_j\big)>I\big(y; x_j|x_k\big)$ does not necessarily mean$I\big(y; x_j|x_k\big)<I\big(y; x_k|x_j\big)$ when two different candidate features x_k and $x_j$are evaluated. This finding implies that maximizing new classification information does not guarantee minimizing redundancy. In light of the above analysis, ICI is introduced in the next section. ICI assembles redundancy information and new classification information into one term. Thus, both evaluation criteria play critical roles simultaneously in finding highly predictive as well as lowly redundant features.

III. Independent Classification Information (ICI)

The major drawbacks faced in the Feature Maximizing Equation (2) are as follows:

1. An inevitable problem is that joint probabilities in Eq. (2) are complicated to be estimated accurately, unless all of the involved variables are independent identically distributed [8].
2. This issue becomes more intractable on small samples in high dimensions.
3. Even if these joint probabilities can be obtained, an exhaustive search of selecting k optimal features from d candidates is near$O(d^k)$, which is almost impractical for high-dimensional learning tasks [9].

A new mutual information term, namely, independent classification information, is defined in this paper. It encompasses both the independent information that a candidate feature provides and the independent information that the selected features preserve. Independent classification information is proved as a loose upper bound of the total classification information of feature subset. Thus, the maximization of independent classification information helps enhance the global discriminative performance. Then, a new feature evaluation criterion, i.e., MRI, is proposed on the basis of independent classification information. Besides pursuing the maximization of feature relevance with classes, MRI maximizes independent classification information. By analysis and comparison with some popular evaluation criteria, MRI is illustrated to properly regulate the effects of feature relevance and feature redundancy, neither of which is exaggerated or depreciated in estimating the contribution of feature to classification. Comprehensive experiments on various data sets testify the effectiveness of MRI in selecting highly predictive and lowly redundant features.Suppose features $x_1, and x_2$ are involved in recognizing the target class y. Then, their independent classification information is defined as

$$ICI(y; x_1, x_2) = I(y; x_1|x_2) + I(y; x_2|x_1) \quad (10)$$

ICI focuses on the amount the specific classification information provided by a feature when feature is given. Suppose one feature is a candidate and the other feature is selected, ICI indicates m, the amount of the new classification information provided by the candidate feature and the amount of the classification information preserved by the selected feature. Mutual information between feature and class and between feature and feature should be further investigated to understand what is measured by ICI.

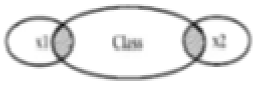


Figure 1: ICI of two statistically independent features

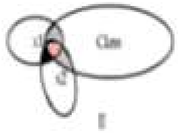


Figure 2: ICI of two partially dependent features

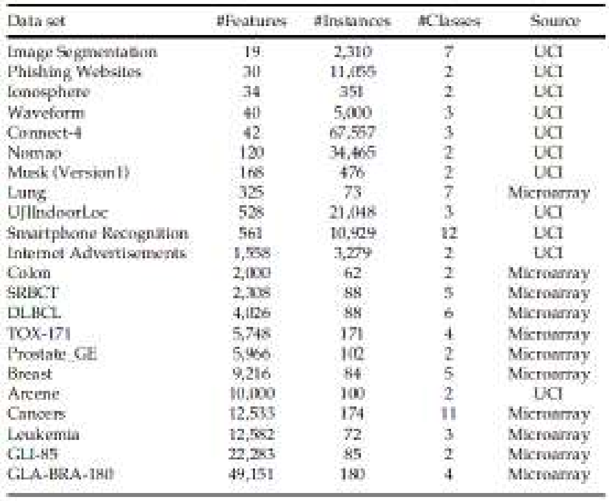
In Figure 1, two features, namely, x1 and x2, are statistically independent from each other, i.e., $p(x1,x2)$=$p(x1)p(x2)$**.** Their classification information is not correlated with each other, i.e., $I(y;x_1|x_2)$=$I(y;x_1)$ and $I(y;x_2|x_1)$ =$I(y;x_2)$. That is, their information for predicting classes is exactly the summation of their respective mutual information with classes. In this case, $ICI(y;x_1|x_2)$=$I(y;x_1)$ + $I(y;x_2)$. Whereas in Figure 2, two features tightly or loosely correlate with each other, which is common in feature selection. The total classification information is provided by two features can be separated to two parts, namely, ICI and dependent classification information. ICI represents the unshared information and comprises two terms, namely,$I(y;x_1|x_2)$and $I(y;x_2|x_1)$. Each term represents the different predictive information of one feature from another feature. Hence, both terms provided respectively by each feature are distinct and helpful for recognizing the target class. They are asymmetric, and cannot be replaced by each other. Another information is the dependent one, which is depicted as the red point part in Fig. 2. This information is the same as that shared by two features. From another angle, this information is the interaction of two features with the target class, which is exactly $I(y;x_1;x_2)$, i.e., the class-relevant redundancy provided by one feature if another feature is selected. In other words, this information fails to help enhance the predictive ability of a subset when a candidate feature is added. The overlapped area in case II also includes a part unrelated to classification, which is exactly$I(x1;x_2|y)$ and marked black in Fig. 2. This information is also a part of the relevance of two features, and is counted as feature redundancy by some selection methods. In fact, this part positively contributes to the joint predictive ability of two features, because large $I(x_1;x_2|y)$ means small$I(y;x_1;x_2)$.Therefore, directly employing the mutual information of two features as their redundancy cannot reflect their actual relationship in classification. One feature redundant with another feature fails to indicate that both features preserve little different classification information.

IV.ExperimentAnd Analaysis

A. Comparing Classification Performance with Non-Mutual- Information Based Feature Selection Approaches

The experiment is to test the classification performance of selected features of the above mentioned benchmark data set, by constructing two classifiers 1- Nearest neighbor (1-NN) classifier and Support Vector Machine(SVM) classifier with 10 fold Cross- validations. The benchmark data sets cover both binary-class and multi-class, and the number of original features varies from less than 50 to near to 50,000. The number of selected features, i.e., k, sequentially increases from 1 to 50 in the interval of 1. That is, the compared criteria respectively select 50 groups of feature subsets whose sizes increase from 1 to 50 for comparison. Two classifiers are constructed for the selected features in the WEKA environment , i.e., 1-Nearest Neighbor (1-NN) classifier and Support Vector Machine (SVM) classifier, andtested with 10-fold cross-validations. Average classification accuracies of both classifiers across the 50 groups of feature subsets selected by each criterion will be recorded. Furthermore, a pairwise t-test at 5 percent significance level will be conducted to evaluate the statistical significance of the results The average accuracies across all the benchmark data sets will be recorded. The filter selection strategies adopted here exclude induction algorithms in selection process thus making thus improving the performance as it becomes independent of the choice of classifiers.

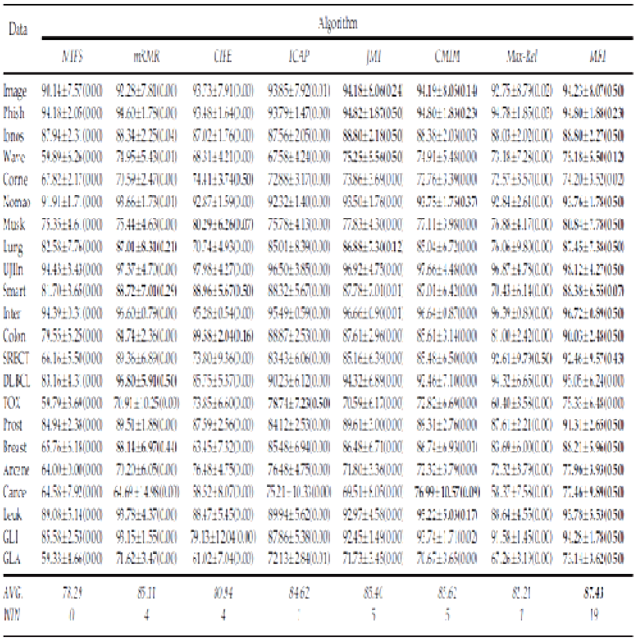
Table 1 : Benchmark Data Sets



Data set	#Features	#Instances	#Classes	Source
Image Segmentation	19	2,310	7	UCI
Phishing Websites	30	11,055	2	UCI
Ionosphere	34	351	2	UCI
Waveform	40	5,000	3	UCI
Connect-4	42	67,557	3	UCI
Nomao	120	34,465	2	UCI
Musk (Version1)	168	476	2	UCI
Lung	325	73	7	Microarray
UJIIndoorLoc	528	21,048	3	UCI
Smartphone Recognition	561	10,929	12	UCI
Internet Advertisements	1,558	3,279	2	UCI
Colon	2,000	62	2	Microarray
SRBCT	2,308	88	5	Microarray
DLBCL	4,026	88	6	Microarray
TOX-171	5,748	171	4	Microarray
Prostate_GE	5,966	102	2	Microarray
Breast	9,216	84	5	Microarray
Arcene	10,000	100	2	UCI
Cancers	12,533	174	11	Microarray
Leukemia	12,582	72	3	Microarray
GLI-85	22,283	85	2	Microarray
GLA-BRA-180	49,151	180	4	Microarray

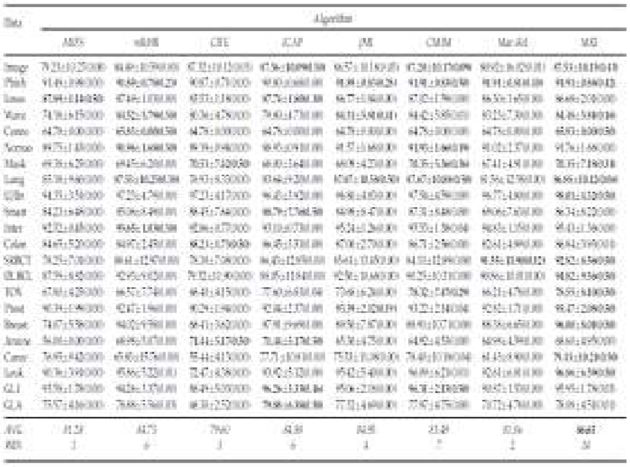
The number of selected features are increased from 5 to 50 in the interval of 5 . Four baseline evaluation criteria, mRMR (minimum Redundancy and Maximum Relevance), CIFE (Conditional Infomax Feature Extraction), JMI(Join Mutual Information), and Max_Rel, are compared with MRI, which are the representative redundancy reduction criteria, new information maximization criterion, and top-k criterion, respectivelyOther metrics, i.e., Balanced Error Rate (BER), Area Under ROC Curve (AUC), Kuncheva's Stability Index (Stability) [30], and Inconsistency Rate [35], are also employed to evaluate the performance of feature subsets. The size of feature subsets increases from 5 to 50 in the interval of 5, and the average BER, AUC, stability, and inconsistency rate across all of the benchmark data sets. Thus, Max-Rel performs best among all of the compared mutual information-based criteria, and is also better than MRI that alleviates feature redundancy in the selected subset. Generally, JMI and CMIM also show comparably better than the other criteria except MRI. That is, these two criteria also have excellent selection abilities.

Table 2: Average 1-NN Classification Accuracy (MeanStd.) with p-Value (in Percentage)



Data	Algorithm							
	MIFS	mRMR	CIFE	ICAP	JMI	CMIM	Max-Rel	MRI
Image	90.14±7.57(0.00)	92.28±7.81(0.00)	93.73±7.91(0.00)	9385±7.92(0.11)	94.18±6.06(0.24)	94.19±6.05(0.14)	92.75±8.79(0.02)	94.23±6.07(0.50)
Phish	94.18±2.06(0.00)	94.60±1.79(0.00)	93.48±1.64(0.00)	9379±1.47(0.00)	94.82±1.85(0.50)	94.80±1.80(0.23)	94.78±1.85(0.03)	94.80±1.80(0.23)
Ionos	87.94±2.3 (0.00)	88.34±2.25(0.04)	87.02±1.76(0.00)	8756±2.05(0.00)	88.80±2.18(0.50)	88.38±2.03(0.03)	88.03±2.02(0.00)	88.80±2.27(0.50)
Wave	59.89±5.26(0.00)	71.95±5.45(0.01)	68.31±4.21(0.00)	6758±4.24(0.00)	7525±5.54(0.50)	74.91±5.48(0.00)	73.18±7.28(0.00)	75.18±6.50(0.12)
Conne	67.82±2.17(0.00)	71.59±2.47(0.00)	74.41±3.74(0.50)	7288±3.17(0.00)	7386±3.69(0.00)	72.76±3.39(0.00)	72.57±3.57(0.00)	74.20±3.52(0.02)
Nomao	91.91±1.7 (0.00)	93.66±1.73(0.01)	92.87±1.59(0.00)	9232±1.40(0.00)	9350±1.76(0.00)	93.73±1.73(0.37)	92.84±2.61(0.00)	93.76±1.78(0.50)
Musk	75.35±4.6 (0.00)	75.44±4.63(0.00)	80.29±6.26(0.07)	7578±4.13(0.00)	77.83±4.30(0.00)	77.11±3.98(0.00)	76.88±4.17(0.00)	80.84±7.78(0.50)
Lung	82.58±7.76(0.00)	87.01±8.30(0.21)	70.74±4.93(0.00)	8501±8.39(0.00)	86.88±7.30(0.12)	85.04±6.72(0.00)	76.06±9.83(0.00)	87.45±7.38(0.50)
UJIIn	94.43±3.43(0.00)	97.37±4.73(0.00)	97.98±4.27(0.00)	9650±3.85(0.00)	96.92±4.73(0.00)	97.66±4.48(0.00)	96.87±4.78(0.00)	98.12±4.27(0.50)
Smart	81.70±3.65(0.00)	88.72±7.00(0.29)	88.96±5.67(0.50)	8832±5.67(0.00)	8778±7.01(0.01)	85.01±6.42(0.00)	70.43±6.14(0.00)	88.38±6.55(0.07)
Inter	94.39±1.3 (0.00)	95.60±0.79(0.00)	95.28±0.54(0.00)	9549±0.59(0.00)	96.66±1.90(0.01)	96.64±0.87(0.00)	96.39±0.83(0.00)	96.72±0.89(0.50)
Colon	79.55±5.25(0.00)	84.74±2.36(0.00)	89.38±2.04(0.16)	8887±2.53(0.00)	87.61±2.96(0.00)	85.61±3.14(0.00)	81.00±2.42(0.00)	90.03±2.48(0.50)
SRBCT	66.16±5.50(0.00)	89.36±6.89(0.00)	73.80±9.36(0.00)	8343±6.06(0.00)	85.16±6.39(0.00)	85.48±6.50(0.00)	92.61±9.79(0.50)	92.46±9.57(0.43)
DLBCL	83.16±4.3 (0.00)	96.80±5.90(0.50)	85.75±5.37(0.00)	9023±6.12(0.00)	94.32±6.89(0.00)	92.46±7.10(0.00)	94.32±6.65(0.00)	95.05±6.24(0.00)
TOX	54.79±3.64(0.00)	70.91±10.25(0.00)	73.85±6.60(0.00)	7874±7.22(0.50)	70.59±6.17(0.00)	72.82±6.69(0.00)	60.40±3.58(0.00)	75.33±6.48(0.00)
Prost	84.94±2.38(0.00)	89.51±1.88(0.00)	87.39±2.56(0.00)	8412±2.53(0.00)	89.61±3.00(0.00)	88.31±2.76(0.00)	87.61±2.21(0.00)	91.31±2.66(0.50)
Breast	65.76±5.18(0.00)	88.14±6.97(0.44)	63.45±7.32(0.00)	8548±6.94(0.00)	86.48±6.71(0.00)	88.74±6.93(0.01)	83.69±6.00(0.00)	88.21±5.96(0.50)
Arcene	64.00±3.00(0.00)	71.23±6.05(0.00)	76.48±4.75(0.00)	7648±4.75(0.00)	71.80±3.36(0.00)	72.32±3.79(0.00)	72.32±3.79(0.00)	77.96±3.93(0.50)
Cance	64.58±7.92(0.00)	64.69±14.98(0.00)	58.52±8.07(0.00)	7521±10.31(0.00)	69.51±6.05(0.00)	76.99±10.57(0.09)	58.37±7.58(0.00)	77.46±9.89(0.50)
Leuk	89.08±5.14(0.00)	93.78±4.37(0.00)	88.47±5.45(0.00)	8994±5.62(0.00)	92.97±4.58(0.00)	95.22±5.03(0.17)	88.64±4.55(0.00)	95.78±5.53(0.50)
GLI	85.58±2.53(0.00)	93.15±1.55(0.00)	79.13±12.04(0.00)	8786±5.38(0.00)	92.45±1.49(0.00)	93.74±1.71(0.02)	91.58±1.45(0.00)	94.28±1.78(0.50)
GLA	59.33±4.66(0.00)	71.62±3.47(0.00)	61.02±7.04(0.00)	7213±2.84(0.01)	71.73±3.45(0.00)	70.67±3.68(0.00)	67.26±3.13(0.00)	73.14±3.62(0.50)
AVG.	78.29	85.11	80.54	84.62	85.40	85.62	82.21	87.43
W/N	0	4	4	1	5	5	1	19

Table 3: Average SVM Classification Accuracy (MeanStd.) with p-Value (in Percentage)



Data	Algorithm							
	MIFS	mRMR	CIFE	ICAP	JMI	CMIM	Max-Rel	MRI
Image	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Phish	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Ionos	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Wave	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Conne	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Nomao	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Musk	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Lung	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
UJIIn	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Smart	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Inter	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Colon	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
SRBCT	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
DLBCL	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
TOX	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Prost	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Breast	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Arcene	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Cance	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Leuk	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
GLI	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
GLA	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
AVG.	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
W/N	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

Generally speaking, it follows from Tables 2 and 3 that MRI is comparable or superior to the other mutual information-based criteria. mRMR, JMI, and CMIM also

perform well, although not better than MRI. The number of selected featuresincreases from 5 to 50 in the interval of 5 (on the datasets of Waveform and Connect, it reaches up to 40). Fourbaseline evaluation criteria, mRMR, CIFE, JMI, and Max_Rel, are compared with MRI, which are the representative redundancy reduction criteria, new information maximization criterion, and top-k criterion, respectively.

V.Conclusion

We have performed a detailed study of a new mutual information term, namely, independent classificationinformation (ICI). It encompassesboth the independent information that a candidatefeature provides and the independent information that theselected features preserve. Independent classification informationis proved as a loose upper bound of the total classificationinformation of feature subset. From the experiments done it clearly shows that the maximizationof independent classification information helps to enhance the overall discriminative performance. Also, a new feature evaluationcriterion, i.e., MRI, is proposed on the basis of independentclassification information. The experiment results show that MRI maximizesindependent classification information. Analysis is done by comparing with some popular evaluation criteria, MRIillustrates in minimizing and regulating effects of feature relevanceand feature redundancy. To concludetheseexperiments on variousdata sets validate the effectiveness of MRI in selecting highlypredictive and lowly redundant features.

References

[1] T. M. Cover and J. A. Thomas, Elements of Information Theory. New

York, NY, USA: Wiley, 1991.

[2] N. X. Vinh, J. Epps, and J. Bailey, “Information theoretic measures for clustering comparison: Variants, properties, normalization and correction for chance,” J.Mach. Learn. Res., vol. 11, pp. 2837–2854, 2010.

[3] I. Guyon and A. Elisseeff, “An introduction to variable and feature selection,” J. Mach. Learn. Res., vol. 3, pp. 1157–1182, 2003.

[4] N. X. Vinh, J. Chan, S. Romano, and J. Bailey, “Effective global approaches for mutual information based feature selection,” in Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining,2014, pp. 512–521.

[5] A. L. Blum and P. Langley, “Selection of relevant features and examples in machine learning,” Artif. Intell., vol. 97, no. 1, pp. 245–271, 1997.

[6] R. Kohavi and G. H. John, “Wrappers for feature subset selection,” Artif.Intell., vol. 97, no. 1, pp. 273–324, 1997.

[7] I. Guyon, S. Gunn, M. Nikravesh, and L. A. Zadeh, Feature Extraction: Foundations and Applications. Berlin, Germany: Springer- Verlag, 2006, ch. 6.

[8] L. Breiman, “Probability,” in Classics in Applied Mathematics, vol. 7. Philadelphia, PA, USA: SIAM, 1992.

[9] L. Yu and H. Liu, “Efficient feature selection via analysis of relevanceand redundancy,” J. Mach. Learn. Res., vol. 5, pp. 1205–1224, 2004.

[10] H. Liu and H. Motoda, Feature Selection for Knowledge Discovery and Data Mining. Boston, MA, USA: Kluwer, 1998.

[11] R. Battiti, “Using mutual information for selecting features in supervised neural net learning,” IEEE Trans. Neural Netw., vol. 5, no. 4, pp. 537–550, Jul. 1994.

[12] H. Peng, F. Long, and C. Ding, “Feature selection based on mutual information criteria of max-dependency, max-relevance, and minredundancy,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 27, no. 8, pp. 1226–1238, Aug. 2005.

[13] P. A. Est_evez, M. Tesmer, C. A. Perez, and J. M. Zurada, “Normalized mutual information feature selection,” IEEE Trans.NeuralNetw., vol. 20, no. 2, pp. 189–201, Feb. 2009.

[14] D. Lin and X. Tang, “Conditional infomax learning: An integratedframework for feature extraction and fusion,” in Proc. 9th Eur.Conf. Comput. Vis., 2006, pp. 68–82.

[15] N. Kwak and C. H. Choi, “Input feature selection for classificationproblems,” IEEE Trans. Neural Netw., vol. 13, no. 1, pp. 143–159,Jan. 2002.

[16] H. Cheng, Z. Qin, C. Feng, Y. Wang, and F. Li, “Conditionalmutual information-based feature selection analyzing for synergyand redundancy,” Electron.Telecommun.Res. Inst. J., vol. 33, no. 2,pp. 210–218, 2011.

[17] A. Jakulin, “Machine learning based on attribute interactions,”Ph.D. dissertation, Faculty Comput. Inf. Sci., Ljubljana Univ.,Ljubljana, Slovenia, 2005.

[18] G. Bontempi and P. E. Meyer, “Causal filter selection in microarraydata,” in Proc. 27th Int. Conf. Mach. Learn., 2010, pp. 95–102.

[19] A. E. Akadi, A. E. Ouardighi, and D. Aboutajdine, “A powerfulfeature selection approach based on mutual information,” Int.J. Comput. Sci. Netw. Secur., vol. 8, no. 4, pp. 116–121, 2008.

[20] R. W. Yeung, “A new outlook on Shannon’s information measures,”IEEE Trans. Inf. Theory, vol. 37, no. 3, pp. 466–474, May 1991.

[21] J. R. Vergara and P. A. Est_evez, "A review of feature selectionmethods based on mutual information," Neural Comput. Appl.,vol. 24, no. 1, pp. 175–186, 2014.

[22] H. Yang and J. Moody, "Data visualization and feature selection:New algorithms for nongaussian data," Advances Neural Inf. Process.Syst., vol. 12, pp. 687–693, 1999.

[23] M. Vidal-Naquet and S. Ullman, "Object recognition with informativefeatures and linear classification," in Proc. 9th IEEE Int.Conf.Comput. Vis., 2003, pp. 281–288.

[24] P. E. Meyer and G. Bontempi, "On the use of variable complementarityfor feature selection in cancer classification," in Applicationsof Evolutionary Computing. Berlin, Germany: Springer,2006, pp. 91–102.

[25] F. Fleuret, "Fast binary feature selection with conditional mutualinformation," J. Mach. Learn.Res., vol. 5, pp. 1531–1555, 2004.

[26] G. Brown, A. Pocock, M. Zhao, and M. Luj_an, "Conditional likelihoodmaximisation: A unifying framework for information theoreticfeature selection," J. Mach. Learn. Res., vol. 13, pp. 27–66, 2012.

[27] K. Bache and M. Lichman, "UCI machine learning repository,"Univ. of California, School Inf. Comput. Sci., Irvine, 2013.[Online]. Available: http://archive.ics.uci.edu/ml

[28] H. Peng, "Mutual information computation," 2007. [Online].Available: http://www.mathworks.com/matlabcentral/fileexchange/14888-mutual-information-computation

[29] J. Tang, S. Alelyani, and H. Liu, "Feature felection for classification:A review," in Data Classification: Algorithms and Applications.Chapman, CA, USA: CRC Press, 2014.

[30] X. He, D. Cai, and P. Niyogi, "Laplacian score for feature selection,"in Proc. Advances Neural Inf. Process. Syst., 2005, pp. 507–514.

[31] G. Roffo, S. Melzi, and M. Cristani, "Infinite feature selection," inProc. IEEE Int. Conf. Comput. Vis., 2015, pp. 4202–4210.

[32] M. Robnik-_Sikonja and I. Kononenko, "Theoretical and empiricalanalysis of ReliefF and RReliefF," Mach. Learn., vol. 53, no. 1/2,pp. 23–69, 2003.

[33] N. Hoque, D. K. Bhattacharyya, and J. K. Kalita, "MIFS-ND: A mutual information-based feature selection method," Expert Syst. Appl., vol. 41, pp. 6371–6385, 2014.

[34] J. M. Sotoca and F. Pla, "Supervised feature selection by clusteringusing conditional mutual information-based distances," PatternRecognition., vol. 43, no. 6, pp. 2068–2081, 2010

[35] M. Dash and H. Liu, "Consistency-based search in feature selection," Artif.Intell., vol. 151, no. 1, pp. 155–176, 2003.