RJMIM: A New Feature Selection Method Based On Joint Mutual Information

Liwen Peng, YongGuo Liu

Knowledge and Data Engineering Laboratory of Chinese Medicine, School of Information and Software Engineering, University of Electronic Science and Technology of China, Chengdu 610054, PR China

Abstract

Feature selection based on information theory plays an important role in classification algorithm due to its computational efficiency and independent from classification method. It is widely used in many application areas like data mining, bioinformatics and machine learning. But drawbacks of these methods are the neglect of the feature interaction and overestimation of features significance due to the limitations of goal functions criterion. To address this problem, we proposed a new feature goal function RJMIM. The method employed joint mutual information and information interaction, which alleviates the shortcomings of overestimation of the feature significance as demonstrated both theoretically and experimentally. The experiments conducted to verify the performance of the proposed method, it compared with four well-known feature selection methods use three publicly available datasets from UCI. The average classification accuracy and C4.5 classifier is used to assess the effectiveness of RJMIM method.

Keywords: feature selection, mutual information, joint mutual information, information Interaction, classification.

1. INTRODUCTION

Feature selection plays an important role in classification algorithms. Feature selection techniques have been widely used in data mining, machine learning and many other fields with the development of specialization data acquisition. Since the relation of features and class are often difficult to measure, many feature-selection techniques were proposed in the past. They can be categorized into three groups: filter (Zhu and Ong, 2007), wrapper (Kohavi and John, 1997) and embedded (Guyon et al., 2003) methods. The wrapper methods are dependent with a classifier usually. So, they have higher computational complexity and have a risk of over-fitting to the classification algorithm. The filter approach selects a subset of features from original datasets with an evaluation criterion. The embedded methods are also dependent with specific classifier, but they are not the same as wrapper approach. These methods use feature subset as part of the training stage. In generally, filter methods are usually faster than wrapper-based methods and embedded methods, and the classification accuracy also good. At present, many researchers proposed evaluate features or feature subsets criteria in the literature, such as Pearson correlation coefficient, distance analysis, mutual information and so on.

Currently, one important factor that affects the classification algorithm performance is the feature evaluation criterion. Mutual information measure based on Information theory is not make an assumption of the linear or nonlinear relationship between the variables, and can process the numerical or categorical data. Because of the advantages of mutual information, it has been widely used in filter feature selection methods. There are some other several MI-based evaluation criterion used to compute the features correlation, conditional mutual information, joint mutual information, mutual information,
and interaction information (Peng 2005; Battiti 1994; Brycki 2008; Meyer 2006; Yihua Lan et al., 2016). Though these existing evaluation criterion based on mutual information have performed well under many circumstances, the drawback of these methods is overestimated some features significance and neglected the interaction between features and the class.

In this paper, we proposed a new feature selection method based on mutual information objective to avoid the limitations of mutual information feature selection approach such as overestimated the feature significance. On the basis of the novel goal function, a sequential forward feature selection algorithm called RJMIM is devised by greedy idea. To verify the performance of the algorithm, the method is compared with four pre-existing feature selection algorithms (JMI, JMIM, IGFS, DISR). Experimental results on three datasets from the UCI database, and it is proves the effectiveness of our algorithm.

The rest of this paper is organized as follow. Section 2 represented some preliminary of entropy and mutual information theory. Section 3 reviews related MI-based feature selection work. Section 4 presents the proposed algorithm. Experiments results are shown in Section 5. Finally, the conclusions are drawn in section 6.

2. BACKGROUND THEORY

This section introduces the theory of information by entropy and mutual information and interpretation the reasons for using them in feature selection. The entropy of a random variable or the mutual information of two variables is measure of its uncertainty (Swingle, 2012). Suppose $X \in \{x_1, ..., x_n\}$ is a discrete random variable, $x_i$ is the possible values that $X$ can take. $C = \{c_n, ..., c_M\}$ is a class label. The entropy of $X$ is then defined as:

$$H(X) = \sum_{i=1}^{N} p(x_i) \log(p(x_i))$$  \hspace{1cm} (1)

where $0 \leq H(X) \leq 1$.

For any two discrete random variables $X$ and $C$, the joint entropy is defined as:

$$H(X,C) = \sum_{i=1}^{N} \sum_{j=1}^{M} p(x_i, c_j) \log(p(x_i, c_j))$$  \hspace{1cm} (2)

Where $p(x_i, c_j)$ is the joint probability of the random variables $X$ and $C$. The $H(X,C)$ is less than or equal to the entropy of both variables.

If one of the two variable is known but the other is not, the remaining uncertainty of variables is computed by the conditional entropy. The conditional entropy is defined as:

$$H(C|X) = \sum_{i=1}^{N} \sum_{j=1}^{M} p(x_i, c_j) \log(p(c_j|x_i))$$  \hspace{1cm} (3)

Where $H(C|X) \leq H(C)$. If the two variables are independent, the conditional entropy is equal to the entropy.

The relation between joint and the conditional entropy is as follow:

$$H(X,C) = H(X) + H(C|X)$$  \hspace{1cm} (4)

The amount of information of both variables share is defined as Mutual Information (Jakulin, 2005). So, the MI defined as follow:
\[ I(X;C) = \sum_{j=1}^{N} \sum_{i=1}^{M} p(x_i, c_j) \log \left( \frac{p(x_i, c_j)}{p(x_i)p(c_j)} \right) \] (5)

Where the value of MI is always positive in equation (5). If the \( I(X;C) \) is high, both of the random variables are strongly related. If the \( I(X;C) \) is zero, both of the random variables are statistically independent. MI can define as the following entropies functions:

\[ I(X;C) = I(X) - H(X|C) \] (6)

The mutual information can used to find the relation between variables \( X \), \( C \) and class \( C \). The feature that has the highest mutual information could consider as the most informative about class \( C \).

Let \( Y \) be a discrete random variable. The Joint mutual information of three variables \( X \), \( Y \) and \( C \), which is defined as follows:

\[ I(X,Y|C) = H(X;C|Y) + I(C;Y) \] (7)

The conditional MI means that the information of two variables in the context of the third variable. (McGill, 1954) proposed the extension of MI, the Multi-information, which allows measuring the MI more than two variables. The multi-information is defined as:

\[ I(X;Y;C) = I(X,Y;C) - I(X;C) - I(Y;C) \] (8)

Currently, the multi-information has not been widely used in literatures. However, some paper (Zhao and Liu, 2009) about more than two variables interaction information used this concept. The variables interaction information are different. Interaction information can be positive, negative or zero (Jakulin, 2005). The value of the Multi-information is zero means that the random variables are independent in the context of the class label (Jakulin, 2003); The value of the Multi-information is negative means that the variables have redundant information; At last, the value of the Multi-information is positive indicated that the random variables together can provide information more than each of them individually.

### 3. RELATED WORK

There are many feature selection algorithms (Frohlich 2003; Lin 2012; Peng 2005; Battiti 1994) based on information theory have been proposed in current. In fact, the problem of feature selection is to find the subset of minimum in the whole set of features with respect to class \( C \). The focus of the work presented in this article is on the filter feature selection methods because of their advantages. Information Gain (Guyon et al., 2003) is the simplest of feature selection methods. It ranks the variables by computing the mutual information between feature and class. Simplicity and low computational cost are the IG algorithms primary advantages. Nevertheless, it does not take into account the interactions of features, the features are not statistically independency in many case. Consequently, some of the selection methods selected features by used redundant information.

(Battiti, 1994) proposed the Mutual Information Feature Selection (MIFS) that attempt to pursue the ‘max-relevancy min-redundancy’ goal. MIFS algorithm select features by adding one by one based on the following criterion:

\[ J_{mifs} = I(X_i;C) - \beta \sum_{X_S \in S} I(X_i;X_S) \] (9)
Where \( S \) is the subset of composed by selected features. The \( \beta \) in the FIMS criterion is a user-defined parameter. It is usually set to \( 0.5 \leq \beta \leq 1 \) by experimentally. If \( \beta = 0 \) then \( J_{\text{mifs}} \) is equivalent to the \( J_{\text{mim}} \) (Lewis et al., 1992).

(Kwok and Choi, 2002) proposed an improvement of MIFS, called MIFS-U, to make a better estimation of the redundancy term based on the criterion:

\[
\text{MIFS-U} = I(x_i;Y) - \beta \sum_{x_j} \frac{I(x_i;Y)}{H(x_j)} I(x_i;x_j)
\]  

(10)

Unlike MIFS that empirically sets \( \beta \) to be one, another variant method mRMR is proposed by (Peng et al., 2005). The redundancy term use cardinality \( |S| \) of the selected features in subset to avoid the method may select irrelevant features when it growing very larger as the subsets expanded. The Minimum-Redundancy Maximum-Relevance criterion is:

\[
\text{mRMR} = I(x_i;Y) - \frac{1}{|S|} \sum_{x_j \in S} I(x_i;x_j)
\]  

(11)

MIFS and his improvement algorithm aim to reduce feature redundancy. (Yang and Moody, 2000) propose a feature selection method called Joint mutual information, which is an alternative criterion that proposed to maximizes the cumulative summation of Joint mutual information with features of selected subset. The feature selection criterion as follows:

\[
\text{JMI} = \sum_{x_i} I(x_i;X_j;Y)
\]  

(12)

The basic idea of JMI is that we should include candidate features that are complementary to the selected features in the subset and given the class labels.

An interesting class of criteria use a normalisation term on the mutual information to offset the inherent bias toward high arity features (Chen and Lin, 2006). A typical of this is Double Input Symmetrical Relevance (Meyer et al., 2006), a modification of the JMI criterion as follow:

\[
\text{DISR} = \sum_{x_i \in S} \frac{I(x_k,x_i;Y)}{H(x_k,x_i,Y)}
\]  

(13)

Other methods that use the feature interaction have been proposed. (Setchi et al., 2013) propose a method FIM which employ feature interaction. (Akadi and Ouardighi, 2008) propose a method which employ Interaction Gain Based Feature Selection (IGFS). IGFS method selects the candidate features that maximized the following criterion:

\[
\text{IGFS} = I(x_i;Y) + \frac{1}{|S|} \sum_{x_i \in S} I(x_i;x_s;Y)
\]  

(14)

There are also some other algorithms which dependence maximizing Feature Interaction. For example, Interaction Capping (IC) proposed by (Brown, 2012).

For choose the most relevance features when select candidate features, JMIM proposed by Rossitza Setchi. It is employs joint mutual information and the ‘maximum of the minimum’ choose the most relevant features according the follow criterion (Bennasar et al, 2015):

\[
\text{JMIM} = \arg \max_{x_i \in F-S} (\min_{x_s \in S} I(x_i,x_s;Y))
\]  

(15)
Generally, the filter methods are based on the concepts of feature relevance, redundancy and complementarity (Vergara et al., 2015). In practice, MIFS, JMI, MIFS-U, mRMR would neglect of feature interaction and overestimation of the significance of features.

4. PROPOSED FEATURE SELECTION ALGORITHM

Generally speaking, lots of the algorithms based on mutual information use the criteria combination of two elements: the relevancy term and the redundancy term. The methods consider that relevancy and redundancy attempt to maximize the relevancy and minimum the redundancy meanwhile such as MIFS, MIFS-U and mRMR et al. The problem of above methods is that the redundancy term could be very large along with the candidate features increase. In this case, the drawback of the methods may select irrelevant features because they are not redundant, but not because the dependence between features and class.

For this reason, JMI method is selected the most relevant feature to the class label in the context of the subset (Brycki et al., 2008). It takes into consideration the relevance of feature and class when the subset of features were selected. However, the JMI method also overweight the importance of some features. Therefore, the problem is n the variable n improvement of JMI, is

above, we compute the interaction information of features and class. Denote a metric by jointing attributes

The concept of interaction information among features defined as the uncertainly caused by jointing attributes X and S in a cartesian product (Jakulin, 2003). \( I(x_i;x_j|C) \) expression the interaction information of features and class. Denote a metric \( \zeta \) that measure the interaction of features with class.

The value of feature interaction can be zero, positive, negative, and it is symmetrical, i.e., \( I(x_i;x_j|C)=I(x_j;x_i|C)=I(C;x_i|x_j) \). When \( I(x_i;x_j|C)>0 \), it means that variable \( x_i \) and \( x_j \) provide extra information together about \( C \) than what could be expected from the two individual feature. When \( I(x_i;x_j|C)<0 \), the negative value means that the variable \( x_i \) and \( x_j \) have partially or completely redundancy information about class. When \( I(x_i;x_j|C)=0 \), it means that the added feature is irrelevant with respect to the dependency between selected features set \( S \) and the class \( C \) (Chen et al., 2011). According to the example above, we compute the interaction information \( I(x_i;x_j|C) \):

\[
\begin{align*}
I(x_2,x_4;C) &= I(x_2;c)+I(x_4;C|x_2) \\
I(x_2,x_3;C) &= I(x_2;c)+I(x_3;C|x_2) \\
I(x_2,x_3,x_4,C) &= I(x_2,x_4;C) = I(x_2,x_1;C) = 0.311. \text{ Therefore, the problem is selecting among } x_1, x_3 \text{ or } x_4.
\end{align*}
\]
\[ I(x_2; x_3; C) = I(x_2, x_3; C) - I(x_2; C) - I(x_3; C) = -0.311; \]
\[ I(x_2; x_1; C) = I(x_2, x_1; C) - I(x_2; C) - I(x_1; C) = 0; \]
\[ I(x_2; x_4; C) = I(x_2, x_4; C) - I(x_2; C) - I(x_4; C) = -0.204. \]

From the result of \( I(x_i; x_j; C) \), we can see that when \( I(x_2, x_3; C) = I(x_2, x_4; C) = I(x_2, x_1; C) \), but

the interaction information among features and class represent different meaning. And

the value of \( I(x_i; x_j; C) \) were not the same at all. Due to the interaction information

is difficult to calculate and interpretation. Therefore, by using (18) – (20), we can obtain:

\[ I(x_2, x_3; C) - I(x_2, x_4; C) = I(x_2; x_3; C) + I(x_3; C) - I(x_2; x_4; C) - I(x_4; C) \tag{21} \]

Because of \( I(x_2, x_3; C) = I(x_2, x_4; C) \), We get the formula as follow:

\[ I(x_2; x_3; C) - I(x_2; x_4; C) = I(x_4; C) - I(x_3; C) \tag{22} \]

Similarly:

\[ I(x_2; x_1; C) - I(x_2; x_3; C) = I(x_3; C) - I(x_1; C) \tag{23} \]

Thus, When the \( I(x_i, x_j; C) = I(x_k, x_j; C) \) or \( I(x_i, x_j; C) = I(x_k, x_j; C) = I(x_\ldots, x_j; C) \), we take into

account the mutual information of candidate feature and class. So, we proposed a new
criterion:

\[ f_{RUMIM} = \max_{f_i \in F \setminus S} [I(x_i; C) + \min_{x_i \in S} I(x_i, x_s; C)] \tag{24} \]

Where \( I(x_i, x_s; C) = I(x_i; C|x_s) + I(x_s; C) \).

The proposed algorithm employs the following forward greedy search strategy to find the

important features for classification:

1. (Initialisation) Set \( X \leftarrow \) “initial set of \( n \) features”;

\( X \leftarrow \) “empty set”.

2. (Computation of the MI with the output class)

\[ \text{For} \forall x_i \in X, \text{compute } I(x_i; C). \]

3. (Choice of the first feature) Find the feature \( x \) that

maximizes \( I(x_i; C) \); set \( X \leftarrow X \setminus \{x_i\} \); set \( S \leftarrow \{x_i\} \).

4. (Greedy selection) Repeat until \( |S| = k \);

\[ \text{(selection of the next feature) choose the feature } f_i = \max_{x_i \in X \setminus S} [I(x_i; C) + \min_{x_i \in S} I(x_i, x_s; C)]; \]

set \( X \leftarrow X \setminus \{x_i\} \); set \( S \leftarrow S \cup \{x_i\} \);

5. (Output) Output the set \( S \) with the selected features.
5. EXPERIMENT

In this section, we conduct the experiments on the public data to evaluate the effectiveness of the proposed algorithm. Our experiments use Matlab toolbox with WEKA on an Interl Core-i5 2.3 GHz processor with 8G memory.

5.1 Experimental methodology

The proposed method RJMIM is compared with the results of the JMI, JMIM, DISR, and IGFS on three (WDBC, mfeat-pixel and Dermatology) benchmarking datasets from the UCI machine-learning repository. Table 1 shows detailed information on the benchmarking. These datasets have different name, number of instances, number of class, number of features and feature types. For numerical features in the benchmarking datasets are discretized by 1+/− Standard Deviation.

For three benchmarking datasets, the average classification accuracy is tested using 10-fold cross-validation, it was conducted to evaluate the performance of each feature selection method. We employed a C4.5 classifier to evaluating the discriminability of selected features.

Table 1 Datasets used in the experiment

<table>
<thead>
<tr>
<th>No</th>
<th>Data Set</th>
<th>Number of features</th>
<th>Number of classes</th>
<th>Number of instances</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>WDBC</td>
<td>30</td>
<td>2</td>
<td>569</td>
</tr>
<tr>
<td>2</td>
<td>mfeat-pixel</td>
<td>240</td>
<td>10</td>
<td>2000</td>
</tr>
<tr>
<td>3</td>
<td>Dermatology</td>
<td>34</td>
<td>6</td>
<td>358</td>
</tr>
</tbody>
</table>

5.2 Experimental result

Figures 1-3 show the average classification accuracy on the three datasets. For the WDBC dataset, 25 features are selected by using each compared algorithm. As shown in Figure 1, the highest average accuracy achieved by RJMIM and DISR is 89.63%, which is higher than the accuracy of JMIM (89.27%). IGFS and JMI algorithm have the same average classification accuracy is 89.45%. Average accuracy of RJMIM method is better than other algorithm, which means that there are some interesting consider both conditional information and multi-information. The difference between JMI and DISR is the normalization, this is the likely cause of the classification accuracy outperform JMI, DISR introduction of more variance by estimating the normalization entropy.

Figure 2 shows the average accuracy on the Dermatology dataset when 25 features were selected. We can see that the feature selection performance of RJMIM is better than all of other compared algorithms. The best accuracy achieved at 93.01% by using RJMIM method, better than the 92.45% obtained by JMIM, and 92.73% obtained by JMI, IGFS, DISR, respectively. In addition, it is obtained the highest accuracy by using RJMIM method on the minimal feature subset selected, the number of selected features are less than JMIM, IGFS, JMI and DISR, separately. Obviously, RJMIM can selected the higher discriminative power features more quickly than other three algorithms. It is also worth noting that the proposed method RJMIM which both balances the conditional redundancy and information interaction outperform all other compared goal functions.

The mfeat-pixel dataset has 50 features selected. As seen in Fig.3, RJMIM achieved the highest average classification accuracy 79.2% when 44 features selected, which is higher than the accuracy of IGFS (78.5%), JMI (78.5%) when 46 features are selected, and better than the 78.8% obtained by JMIM and DISR (77.3%). As mentioned above section,
when the number of selected features grows, most of the method that use the cumulative sum approximation will be overestimating some features. The RJMIM employs feature interaction and ‘maximum of the minimum’, which should choose the most relevant features with class. The results of experiment validate that this new evaluation criterion is reasonable.

Table 2 summarizes the average classification accuracy achieved in three datasets and respective classification accuracy measured on a subset of 20, 25, 45, 50 features. As seen in table 2, the proposed algorithm (RJMIM) can find out the features with high discriminate power than other four methods. On the WDBC dataset, RJMIM and DISR achieve the highest classification accuracy when 20 and 25 features selected, respectively. RJMIM have better accuracy (91.89% with 20 features and 93.01% with 25 features) on the Dermatology dataset. With 45 features, RJMIM produces best accuracy (79.2%) on the mfeat-pixel dataset. RJMIM also outperforms JMI (78.5%), IGFS (78.5%), DISR (77.35%) and JMIM (78.8%) when 50 features selected. The superiority of RJMIM can be due to the reason that RJMIM consider both of conditional redundancy and information interaction. Moreover, the algorithm does not increase time complexity compared to JMIM when it takes account of two aspects.

**Table 2** Classification accuracy at different feature number

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Average classification accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>WDBC</td>
<td>Dermatology</td>
</tr>
<tr>
<td>Number of features</td>
<td>20</td>
</tr>
<tr>
<td>JMI</td>
<td>89.10%</td>
</tr>
<tr>
<td>IGFS</td>
<td>89.10%</td>
</tr>
<tr>
<td>DISR</td>
<td><strong>89.63%</strong></td>
</tr>
<tr>
<td>JMIM</td>
<td>89.27%</td>
</tr>
<tr>
<td>RJMIM</td>
<td><strong>89.63%</strong></td>
</tr>
</tbody>
</table>

**Figure 1.** Average classification accuracy on WDBC dataset.
6. CONCLUSION

In this paper, we present a new greedy feature selection method based on joint mutual information. The proposed algorithm selects the features with high discriminate power by measure both of the joint mutual information and interaction information between the features already selected and candidate features. The method has been evaluated in terms of average classification accuracy by experiment on three UCI datasets and compared with four other feature selection algorithms: Joint Mutual Information (JMI), Joint Mutual Information Maximisation (JMIM), Double Input Symmetrical Minimum Redundancy (DISR), interaction gain for Feature Selection (IGFS). To evaluate the performance of the proposed algorithm, the average classification accuracy used in selected feature subset that employed C4.5 classifier on three public datasets at different number features. The results shows good performance, which is better than or at the same level as the other four algorithms. It is demonstrates that the ability and reasonable of the proposed method.
Future work includes proposed new criterion based on joint mutual information and using other search strategies to find global mutual information-based feature selection algorithm. In addition, further investigation of the information interaction and improve the classification accuracy can be consider in the future research.

7. ACKNOWLEDGMENTS

This research was supported in part by the National Natural Science Foundation of China (NSFC) under grants 60903074, the National High Technology Research and Development Program of China (863 Program) under grant 2008AA01Z119.

8. REFERENCES