Indonesian Journal of Electrical Engineering and Computer Science
Vol. 22, No. 3, June 2021, pp. 1619∼1628
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v22i3.pp1619-1628

Unsupervised feature selection with least-squares quadratic mutual information

Janya Sainui, Chouvanee Srivisal
Division of Computational Science, Faculty of Science, Prince of Songkla University, Songkhla, Thailand

Article Info

Article history:

Received Dec 27, 2020
Revised May 10, 2021
Accepted May 24, 2021

Keywords:

Least-squares quadratic mutual information
Mutual information
Unsupervised feature selection

ABSTRACT

We propose the feature selection method based on the dependency between features in an unsupervised manner. The underlying assumption is that the most important feature should provide high dependency between itself and the rest of the features. Therefore, the top m features with maximum dependency scores should be selected, but the redundant features should be ignored. To deal with this problem, the objective function that is applied to evaluate the dependency between features plays a crucial role. However, previous methods mainly used the mutual information (MI), where the MI estimator based on the k-nearest neighbor graph, resulting in its estimation dependent on the selection of parameter, k, without a systematic way to select it. This implies that the MI estimator tends to be less reliable. Here, we introduce the least-squares quadratic mutual information (LSQMI) that is more sensible because its tuning parameters can be selected by cross-validation. We show through the experiments that the use of LSQMI performed better than that of MI. In addition, we compared the proposed method to the three counterpart methods using six UCI benchmark datasets. The results demonstrated that the proposed method is useful for selecting the informative features as well as discarding the redundant ones.

Corresponding Author:

Janya Sainui
Division of Computational Science, Faculty of Science
Prince of Songkla University
Songkhla, Thailand
Email: janya.s@psu.ac.th

1. INTRODUCTION

Feature selection aims to select the most informative subset of features to capture the structure of the original data. It usually appears as a pre-processing step for various tasks such as classification [1], [2], clustering [3]-[5], data mining [6], resulting in better results. In a supervised scenario, the idea is to select the features that are the most relevant to the class labels [7], [8]. However, in this paper, we focus on an unsupervised feature selection that is more difficult to achieve than the supervised feature selection due to the absence of the class labels [9]-[14]. In unsupervised manner, feature selection is important as it helps improve the performance as well as reduce the computational time of clustering [15]-[19]. Moreover, unsupervised feature selections may be used for various purposes such as visualization and pre-processing step for supervised learning [7], [15], [20]. Unsupervised feature selection can be divided into two categories, namely wrapper method and filter method [10], [15]. Many existing methods of unsupervised feature selection are based on the wrapper method such that the selected features are dependent on the specific clustering algorithm. The main drawbacks of the wrapper method are its computational complexity and limited use with a particular clustering

algorithm. In this paper, we focus on the filter method which evaluates the important features without using any clustering algorithm [21]-[25]. Thus, the filter method is more general and useful than the wrapper method. We review some existing unsupervised feature selection methods relating to our work as follows.

Laplacian score (LS) [10] is based on Laplacian Eigenmap and locality preserving projection. For each feature, the Laplacian score is estimated using the nearest neighbor graph. If it is a good feature, its LS tends to be small. The LS is based on the observation that, two data points are probably related to the same cluster if they are close to each other. The drawbacks of this method are that there is no systematic way to tune the number of nearest neighbor k or Gaussian width, and the redundant features are not observed. multi-cluster feature selection (MCFS) [11] selects the features while preserving the cluster structure of the data. The authors proposed to use multiple eigenvectors of graph Laplacian, which are defined on the affinity matrix of data points to capture the multi-cluster structure of data. The algorithm performs especially well when the number of selected features is small (e.g., $<= 50$). However, it performs best when the number of used eigenvectors is equal to the number of clusters, but in an unsupervised manner, the number of clusters remains unknown. Moreover, we observed through the experiments that this method tends to be sensitive to noise features. Unsupervised feature selection based on mutual information (UFSMI) [9] is derived from the observation that good features share information in common, and noisy features are less correlated with the other features. Mutual information (MI) is then used as the objective function to capture the shared information between a feature and the rest of features, and they choose the first m features that achieve the higher mutual information. The higher level of noise leads to a smaller average value of the score function because the addition of noise reduces the mutual information between features. The desired property of the feature selection algorithm aims at removing noisy features. However, the estimation of MI is less reliable as it is computed using the k-nearest neighbor graph. Thus, its performance is based on the tuning parameter, k, resulting in that a subset of the selected features may not be the best.

All methods mentioned above including several existing methods such as [12]-[14] are based on the k-nearest neighbor graph or other tuning parameters, and there is no systematic way to choose such parameters (i.e., k). In addition, they did not observe the redundant features. Therefore, in this paper, we would like to deal with these problems. Our idea is inspired by [9] as they proved that a good feature is expected to be dependent on the rest of features. However, the method did not take the redundancy of features into account, and thus may not produce an optimal feature subset. Moreover, in [9], they used the mutual information (MI) as their dependent measure, and the MI is estimated based on the k-nearest neighbors which lacks of the systematic way to choose the appropriate k. So that the estimation of MI tends to be less reliable. In order to solve these problems, we propose to apply the L_2-distance variant of MI called the quadratic mutual information (QMI) [26], where the least-square method is used to approximate QMI [27]. Therefore, the used parameters can be obtained automatically by cross-validation. The contribution of this paper is a novel method of unsupervised feature selection based on the least-squares quadratic mutual information (LSQMI), which is powerful for selecting the most important features as well as rejecting the redundant features. In the experimental results, we show that the LSQMI is more reliable than MI. In addition, we demonstrate that the proposed method is promising through the experimental results on six UCI machine learning repositories.

The rest of paper is organized as follows. In section 2, we review the least-square quadratic mutual information and its normalization. We describe the proposed method in section 3. The experimental setup and results are shown in section 4. In section 5, we provide the conclusion.

2. THE LEAST-SQUARE QUADRATIC MUTUAL INFORMATION

In this section, we first review the least-square quadratic mutual information (LSQMI) [27] that can be used to measure a statistical dependence among features. As LSQMI is ranged from 0 to ∞, we need to normalize LSQMI to [0,1] range. We then review the normalization of LSQMI [28] that is finally used as the objective function for our proposed unsupervised feature selection method.

2.1. LSQMI estimation

Suppose that we are given a set of $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^{n}$ independently drawn from a joint probability distribution with density $p(\boldsymbol{x}, \boldsymbol{y})$. Here, $\boldsymbol{x} \in R^n$ is a featrue vector and $\boldsymbol{y} \in R^{n\times(d-1)}$ is the rest of feature vectors, where n denotes the number of samples and d denotes the number of features. The quadratic mutual information (QMI) [26] between $\boldsymbol{x}$ and $\boldsymbol{y}$ is defined as

$$\mathrm{QMI} := \iint \Big(f(\boldsymbol{x},\boldsymbol{y})\Big)^2 \mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{y},$$

where

$$f(\boldsymbol{x},\boldsymbol{y}) := p(\boldsymbol{x},\boldsymbol{y}) - p(\boldsymbol{x})p(\boldsymbol{y}),$$

and $p(\boldsymbol{x})$ and $p(\boldsymbol{y})$ denote the marginal densities of $\boldsymbol{x}$ and $\boldsymbol{y}$, respectively. In LSQMI [27], the density difference $f(\boldsymbol{x},\boldsymbol{y})$ is modeled as

$$g(\boldsymbol{x},\boldsymbol{y}) = \sum_{\ell=1}^{n} \boldsymbol{\theta}_\ell K(\boldsymbol{x},\boldsymbol{x}_\ell) L(\boldsymbol{y},\boldsymbol{y}_\ell),$$

where $K(\boldsymbol{x},\boldsymbol{x}_\ell)$ and $L(\boldsymbol{y},\boldsymbol{y}_\ell)$ are kernel functions for $\boldsymbol{x}$ and $\boldsymbol{y}$, respectively. Then, $\boldsymbol{\theta} = (\theta_1,\ldots,\theta_n)^\top$ is learned by least-squares as

$$\min_{\boldsymbol{\theta}} \iint \Big(g(\boldsymbol{x},\boldsymbol{y}) - f(\boldsymbol{x},\boldsymbol{y})\Big)^2 \mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{y}.$$

An empirical and regularized version of the above optimization problem is given as

$$\widehat{\boldsymbol{\theta}} := \operatorname*{argmin}_{\boldsymbol{\theta}} \left[\boldsymbol{\theta}^\top \boldsymbol{H}\boldsymbol{\theta} - 2\boldsymbol{\theta}^\top \widehat{\boldsymbol{h}} + \lambda \boldsymbol{\theta}^\top \boldsymbol{\theta}\right],$$

where $\lambda \geq 0$ is the regularization parameter, and $\boldsymbol{H}$ and $\widehat{\boldsymbol{h}}$ are defined as

$$\begin{aligned} H_{\ell,\ell'} &:= \int K(\boldsymbol{x},\boldsymbol{x}_\ell)K(\boldsymbol{x},\boldsymbol{x}_{\ell'})\mathrm{d}\boldsymbol{x} \int L(\boldsymbol{y},\boldsymbol{y}_\ell)L(\boldsymbol{y},\boldsymbol{y}_{\ell'})\mathrm{d}\boldsymbol{y}, \\ \widehat{h}_\ell &:= \frac{1}{n}\sum_{i=1}^{n} K(\boldsymbol{x}_i,\boldsymbol{x}_\ell)L(\boldsymbol{y}_i,\boldsymbol{y}_\ell) - \frac{1}{n^2}\sum_{i,j=1}^{n} K(\boldsymbol{x}_i,\boldsymbol{x}_\ell)L(\boldsymbol{y}_j,\boldsymbol{y}_\ell). \end{aligned} \tag{1}$$

In this paper, we use the Gaussian kernel for both $K(\boldsymbol{x},\boldsymbol{x}_\ell)$ and $L(\boldsymbol{y},\boldsymbol{y}_\ell)$ as both $\boldsymbol{x}$ and $\boldsymbol{y}$ are continuous:

$$\begin{aligned} K(\boldsymbol{x},\boldsymbol{x}_\ell) &:= \exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{x}_\ell\|^2}{2\sigma^2}\right), \\ L(\boldsymbol{y},\boldsymbol{y}_\ell) &:= \exp\left(-\frac{\|\boldsymbol{y}-\boldsymbol{y}_\ell\|^2}{2\sigma^2}\right), \end{aligned}$$

so that the integral in $\boldsymbol{H}$ (1) can be computed analytically as

$$H_{\ell,\ell'} = (\pi\sigma^2)^{d_x/2} \exp\left(-\frac{\|\boldsymbol{x}_\ell - \boldsymbol{x}_\ell\|^2}{4\sigma^2}\right) \times (\pi\sigma^2)^{d_y/2} \exp\left(-\frac{\|\boldsymbol{y}_\ell - \boldsymbol{y}_{\ell'}\|^2}{4\sigma^2}\right).$$

Thus, the solution $\widehat{\boldsymbol{\theta}}$ can be obtained as

$$\widehat{\boldsymbol{\theta}} = (\boldsymbol{H} + \lambda \boldsymbol{I})^{-1}\widehat{\boldsymbol{h}},$$

where $\boldsymbol{I}$ denotes the identity matrix. Finally, the least-squares quadratic mutual information (LSQMI) is given by

$$\widehat{\mathrm{QMI}} := \max(0, 2\widehat{\boldsymbol{\theta}}^\top \widehat{\boldsymbol{h}} - \widehat{\boldsymbol{\theta}}^\top \boldsymbol{H}\widehat{\boldsymbol{\theta}}).$$

2.2. The normalisation of LSQMI

As LSQMI is ranged from 0 to ∞, we here normalize it as follow:

$$\widehat{\mathrm{NQMI}} := \frac{\widehat{\mathrm{QMI}}(\boldsymbol{x}, \boldsymbol{y})}{\max(\widehat{\mathrm{QMI}}(\boldsymbol{x}, \boldsymbol{x}), \widehat{\mathrm{QMI}}(\boldsymbol{y}, \boldsymbol{y}))}. \tag{2}$$

The higher $\widehat{\mathrm{NQMI}}$ indicates the more information shared between the feature $\boldsymbol{x}$ and the rest of features $\boldsymbol{y}$, while it is closed to zero if the feature $\boldsymbol{x}$ is a noise feature. We assume that the values of $\widehat{\mathrm{QMI}}(\boldsymbol{y}, \boldsymbol{y})$ and $\widehat{\mathrm{QMI}}(\boldsymbol{x}, \boldsymbol{x})$ are larger than zero and that of $\widehat{\mathrm{QMI}}(\boldsymbol{x}, \boldsymbol{y})$, because they are the dependency of themselves. Now, the range of $\widehat{\mathrm{NQMI}}$ is $[0, 1]$. However, in practice, $\widehat{\mathrm{NQMI}}$ may be higher than 1.

3. THE PROPOSED UNSUPERVISED FEATURE SELECTION METHOD

In an unsupervised scenario, the objective function that is used to evaluate the important feature is very important. However, the objective function of the existing methods tends to be less reliable because its approximation depends on the parameters included in the objective function, and there is no automatic way to choose those parameters. In this paper, we propose an unsupervised feature selection by using a more reliable objective function called the least-squares quadratic mutual information (LSQMI). The benefits of LSQMI are that the parameters needed in the objective function can be chosen by cross validation, and it is less sensitive to noises and outliers that may cause the estimation of the selected criteria unreliable [26], [27]. The proposed method is described as follows.

Given a set of n samples denoted as $\{\boldsymbol{x}_i\}_{i=1}^{n}$, where each sample $\boldsymbol{x}_i$ consists of d features (i.e., $\boldsymbol{x}_i = \{f_1, \ldots, f_d\}$), our goal here is to select the most m informative features from d features, while the redundant features should be discarded. To address this problem, we apply the LSQMI [27] as our criteria for both selecting the most m important features as well as ignoring the redundant ones. Let $\mathbf{f}_j \in R^n$ be the j^{th} feature, and $\mathbf{f}_{\setminus j} \in R^{n \times (d-1)}$) be all features excluding $\mathbf{f}_j$. To select the important but not redundant features, our proposed algorithm is shown in Algorithm 1.

Algorithm 1 The unsupervised feature selection with least-squares quadratic mutual information (UFSLSQMI)

Input: A set of features, $\{\mathbf{f}_1, ..., \mathbf{f}_d\}$, $\mathbf{f}_j \in R^n$, the number of needed features, m, and a threshold, τ.
Output: A subset of selected features, $S = \{\mathbf{f}_1^*, \ldots, \mathbf{f}_m^*\}$.

$S \leftarrow \varnothing$
for $j = 1, \ldots, d$ **do**
  compute $\widehat{\mathrm{NQMI}}$ (Equation (2)) between $\mathbf{f}_j$ and $\mathbf{f}_{\setminus j}$
end for
sort features $\{\mathbf{f}_1, ..., \mathbf{f}_d\}$ according to their $\widehat{\mathrm{NQMI}}$ in descending order resulting in $\{\mathbf{f}_1^*, ..., \mathbf{f}_d^*\}$
$S \leftarrow \{S \cup \mathbf{f}_1^*\}$
$k \leftarrow 1$
for $j = 2, \ldots, d$ **do**
  if $(k < m)$ **then**
    if $(\frac{\widehat{\mathrm{NQMI}}(\mathbf{f}_j^*, \mathbf{f}_{\setminus j}^*)}{\widehat{\mathrm{NQMI}}(\mathbf{f}_{j-1}^*, \mathbf{f}_{\setminus j-1}^*)} < \tau)$ **then**
      $S \leftarrow \{S \cup \mathbf{f}_j^*\}$
      $k \leftarrow k + 1$
    end if
  else
    break
  end if
end for

The inputs of the algorithm are a set of features, $\{\mathbf{f}_1, ..., \mathbf{f}_d\}$, the number of needed features, m, and a threshold, τ. While the output is a subset of selected features $S = \{\mathbf{f}_1^*, \ldots, \mathbf{f}_m^*\}$. The algorithm starts by

computing the normalisation of LSQMI as (2) for all features (lines 2 - 4). Then, the features are sorted in descending order according to their score (line 5). Lastly, the features are selected according to the ranking; a feature with the highest score is firstly selected and the feature with the second high score is then selected and so on. However, the redundant features are ignored by discarding the feature that has the normalisation of LSQMI score closed to the previous one. In other words, if the ratio of the current score and the previous one is higher than the given threshold τ, (line 10), the feature according to the smaller score is ignored. The algorithm is repeated until the m features (lines 8 - 17) are obtained.

The performance of the proposed method depends on the parameters used for approximating the LSQMI. However, thanks to the least-squares method, we can choose the appropriate parameters by cross-validation method. Although the selected features are also depended on the threshold τ, we show you through the experiments that our objective function well defines the redundant features and the noise features. In other words, if two features are redundancy, their LSQMI scores tend to be closed to each other, and if a feature is the noise feature, the LSQMI score tends to be small.

4. RESULTS AND DISCUSSION

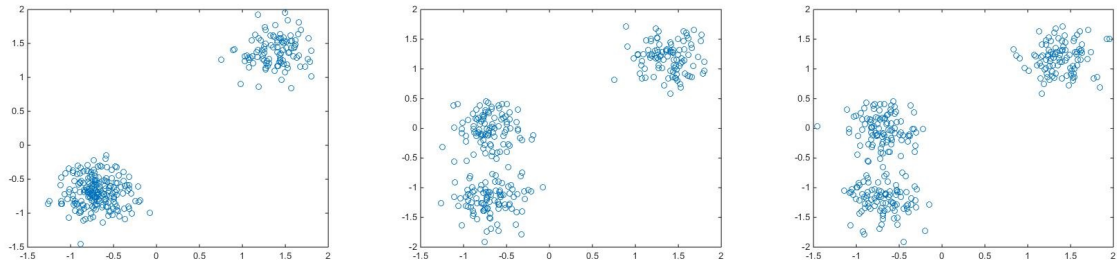
In this section, we demonstrate the effectiveness of the proposed unsupervised feature selection. We compared the proposed method with three competitive methods, including UFSMI [9], LSFS [10], and MCFS [11]. All methods are based on the k-nearest neighbor, and here we set $k = 5$ for all methods as their default setting. For MCFS, there is another parameter that is the number of eigenvectors, and we set this parameter at 5 as the same reason. For the proposed method, the threshold τ is set at 0.95.

The competitive methods including the proposed method rank the features according to the criteria to select the top m features. This means that the objective function mainly affect a subset of selected features. Thanks to the reliability of LSQMI, our method considers not only the importance of features but the redundancy also. Therefore, in the experiment setups, we firstly show you that our objective function is effective for selecting the important features as well as discarding the redundant features on the synthetic data. Secondly, we illustrate the clustering performance after feature selection using six UCI benchmark datasets to confirm the usefulness of our proposed method.

4.1. Synthetic data

In the first experiment, we would like to demonstrate the reliability of the LSQMI comparing to the competitive methods. Specifically, we would like to show you that the LSQMI is more reliable for evaluating both the informative feature and the redundant features. To do so, we conducted a synthetic data of 3 clusters with 3-dimensionality illustrated in Figure 1. We generated 100 instances with standard deviation = 1 for each class. Figure 1 (a)-(c) shows that the 3^{rd} feature is the most important for underlying clusters, while the 1^{st} feature and the 2^{nd} feature are redundancy, because they have the same mean.

Therefore, if we would like to choose two features (i.e., $m = 2$), the first selected one should be the 3^{rd} feature, and the second one may be the 1^{st} feature or the 2^{nd} feature. In other words, the set of selected features should not be the 1^{st} feature and the 2^{nd} feature, because they can capture only 2 clusters of the original data as shown in Figure 1 (a).



(a) (b) (c)

Figure 1. Relationship between features in the synthetic data: (a) Feature 1 vs. Feature 2, (b) Feature 1 vs. Feature 3, (c) Feature 2 vs. Feature 3

We ran all methods using this toy data, where each method was repeated 10 times with different random data. We found that UFSMI performs the worst, because the 3^{rd} feature was not selected 4 times. LS missed selecting the 3^{rd} feature 1 time, while the proposed method and MCFS always selected the 3^{rd} feature. The average scores between each feature and the rest features obtained from each objective function are shown in Table 1, showing that our objective function, $\widehat{\text{NQMI}}$, tends to be more sensible than other ones. As can be seen, the $\widehat{\text{NQMI}}$ score between the 3^{rd} feature and the rest features is significantly highest, while the $\widehat{\text{NQMI}}$ scores according to the 2^{nd} feature and the 1^{st} feature are close to each other. In contrast to other objective functions, the scores according to each feature are almost the same for all features. Notice that the scores of LS showing in Tables 1 and 2 are the scores of (1 − Laplacian score) for each feature. Thus, as same as other objective functions, the feature with larger score is more important.

Table 1. The average scores over 10 runs between the feature j and the rest of the features obtained from different objective functions using the synthetic data

Objective functions	feature 1	feature 2	feature 3
LS	0.9935	0.9934	0.9936
MI	-9.3682	-9.3771	-9.3730
MCFS	0.7330	0.8010	0.9848
$\widehat{\text{NQMI}}$	0.7424	0.7404	1.9034

Table 2. The average scores over 10 runs between the feature j and the rest of the features obtained from different objective functions after adding the noise feature (feature 4)

Objective functions	feature 1	feature 2	feature 3	feature 4
LS	0.9821	0.9818	0.9801	0.9704
MI	-8.1727	-8.1980	-8.2047	-8.8783
MCFS	0.6872	0.7244	0.8226	0.9997
$\widehat{\text{NQMI}}$	0.3331	0.3415	0.8952	0.0167

Next, we added a noise feature (the 4^{th} feature) with mean = 0 for all clusters and standard deviation = 1 into the above toy data. We ran the experiment 10 times for each method again, and the results show that our method outperforms other methods including MCFS. Moreover, MCFS performs the worst in this experiment as it always selected the 4^{th} feature, while the other three methods never selected the 4^{th} feature. The average scores of each method are shown in Table 2, confirming that $\widehat{\text{NQMI}}$ is more reliable than other objective functions. Notice that MCFS works well if the number of used eigenvectors is equal to the number of clusters; however, in an unsupervised scenario, we do not know the actual number of clusters.

4.2. UCI datasets

Here, we evaluated the clustering performance after feature selection using six UCI benchmark datasets, namely Abalone, Sonar, Glass, Pima, Heart, and Cancer datasets. For each dataset, the number of clusters (c), the number of features (d), and the number of samples (n) are shown in Figure 2. For each run, we randomly chosen 90% samples from each dataset, then we performed feature selection. Finally, we performed clustering by the k-means algorithm; we ran the k-means algorithm 10 times with random initialization and chose the best solution with the minimum error rate. The clustering accuracy is evaluated. The results of each dataset are shown in Figure 2, showing the average accuracies over 10 runs at different number of seleced features, m.

The results indicate that the proposed method (i.e., UFSLSQMI) almost performs better than the counterpart approaches. However, we cannot say that the proposed method is the best. Because, in practice, there is no method that works well for all data. Moreover, it is hard to explain the reasons why it works or it does not work. What we can say is that the proposed method can be an alternative for solving unsupervised feature selection problem that may work well for your data.

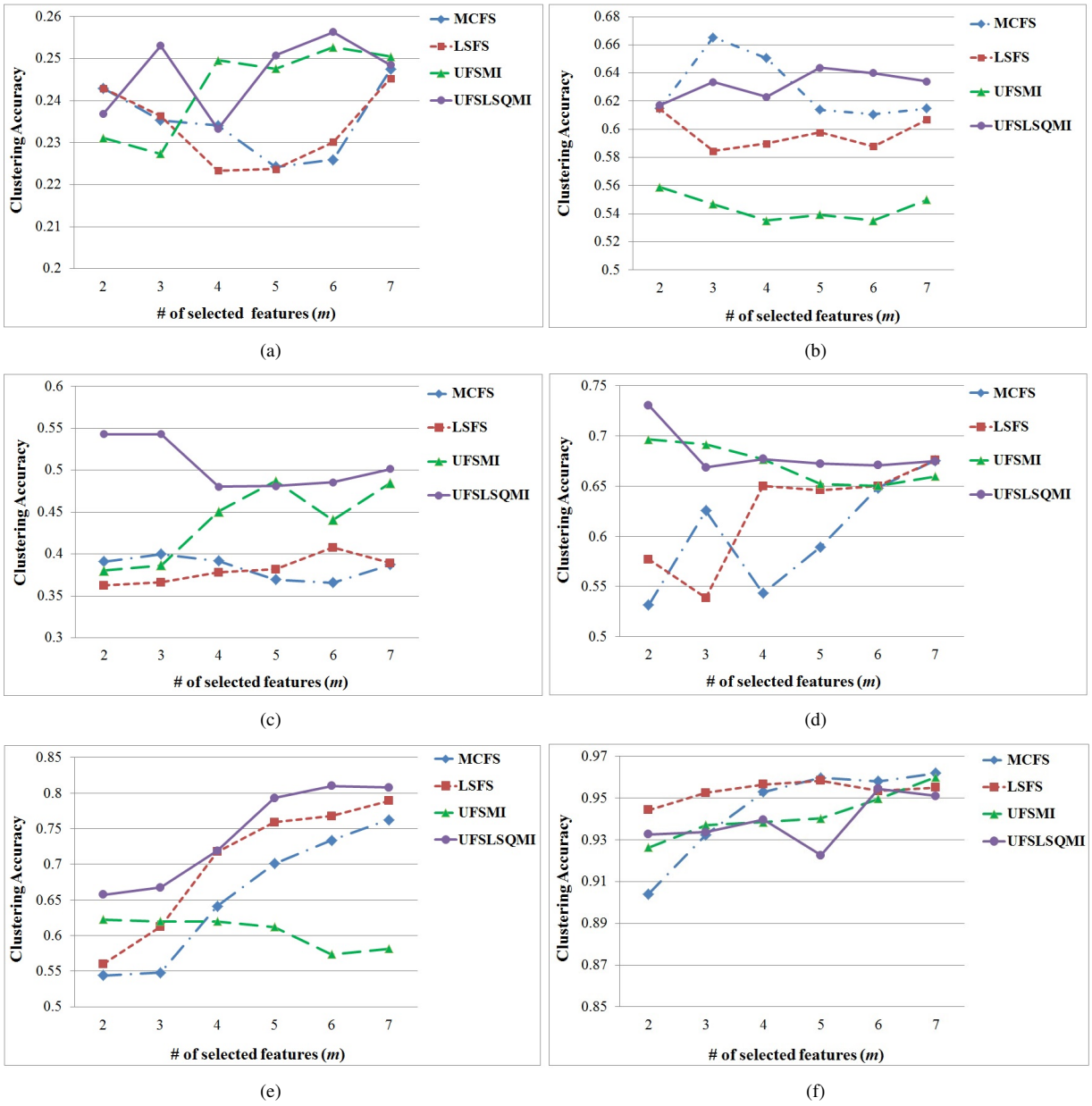


Figure 2. The average of clustering accuracies over 10 runs on UCI datasets: (a) Abalone (c=10 ,d=8 , n=2493), (b) Sonar (c=2 ,d= 60, n=188), (c) Glass (c=6 ,d=9 , n= 193), (d) Pima (c= 2,d= 8, n= 692), (e) Heart (c=2 ,d=13 , n=244), (f) Cancer (c=2 ,d=9 , n=615)

4.3. Discussion

In the experiments, we compared the proposed unsupervised feature selection method with the other filter methods. All methods including the proposed method select the informative features from the rank of features according to the scores of the objective function. This means that the objective function plays an important role for this task. We discuss here each objective function as follows. Firstly, Laplacian score (LS) is computed by constructing of a nearest neighbor graph with the specific-defined number of nearest neighbors, k. This is the reason why the Laplacian score feature selection does not always perform well. In addition, as can be seen in the experiments on toy data (Table 1 and Table 2), the Laplacian scores of the most important feature and the noise feature are not significantly different from each other, this may cause the rank of features relating to the important features is unreliable.

Secondly, the mutual information (MI) is also computed based on the k-nearest neighbors graph. The results in Tables 1 and 2 demonstrate that this MI score does not describe the difference between the informative feature and the noise feature well. In other words, as same as LS, the MI scores of the important feature and the noise feature are not significantly different from each other.

Thirdly, the MCFS score is also computed based on the k-nearest neighbor graph such that the suitable k effects the good score. Not only the parameter k, the MCFS score also needs the number of used eigenvectors that makes the MCFS score depend on these two parameters. As can be seen in the experiment on toy data with noise feature, the MCFS score of noise feature is unreliable; however, if we change the number of used eigenvectors to be 2 instead of 5, the MCFS score of noise feature becomes more reliable. This confirms that the MCFS score heavily depends on the number of used eigenvectors.

Lastly, the proposed method utilizes the LSQMI that approximates by least-square method so that the cross-validation can be included, resulting in that the estimator of QMI is sensible. As can be seen in Tables 1 and 2, the normalized LSQMI describe the structure of the original data as well; the LSQMI is highest if the feature is the most important as the 3^{rd} feature of the toy data; in contrast, the LSQMI is smallest if the feature is the noise feature that provides no information about data like the 4^{th} feature. The main drawback of the proposed method is the time complexity that is higher than other methods because of the cross-validation process.

To sum up, the main problem of the filter based unsupervised feature selection methods is lacking of the ability to automatically choose the parameters included in their objective function so that the approximation of the objective function may be unreliable. Thus, we here propose to use the LSQMI for unsupervised feature selection, that is not only reliable for evaluating the important features but also for ignoring the redundant features.

5. CONCLUSION

The goal of feature selection is to select a subset of features in order to reduce the dimensionality of the original data, while the structure of data should be preserved. There are two main points that are important in order to achieve this goal. Firstly, we need to select the most informative features underlying preserving the structure of the original data. Secondly, we have to remove the redundant features because they do not provide more information about data. However, the existing approaches may not select the important features because the use of unreliable objective functions. Moreover, they do not ignore the redundant features. In this paper, we focus on these two points by exploring the reliable objective function that is the most important for evaluating the informative features as well as the redundant features in an unsupervised manner. Specifically, we propose to use the least-squares quadratic mutual information (LSQMI) as the objective function to select the valuable features excluding the redundant features. The benefit of LSQMI is that the needed parameters can be chosen by cross validation resulting that the estimation of the objective function will be more reliable, while most of the other objective functions lack of the systematic way to choose their parameters. Through the experiments, we show that the LSQMI can be used to determine the important features including the redundant feature as well.

REFERENCES

[1] S. B. Kotsiantis, "Feature selection for machine learning classification problems: a recent overview," *Artificial Intelligence Review*, vol. 42, no. 1, pp. 157–176, 2011, doi: 10.1007/s10462-011-9230-1.

[2] J. Cai, J. Luo, S. Wang, and S. Yang, "Feature selection in machine learning: a new perspective," *Neurocomputing*, vol. 300, pp. 70-79, 2018, doi: 10.1016/j.neucom.2017.11.077.

[3] L. Talavera, C. Nord, and J. Girona, "Dependency-based feature selection for clustering symbolic data," *Intelligent Data Analysis*, vol. 4, no. 1, pp. 19-28, 2000, doi: 10.3233/IDA-2000-4103.

[4] V. Roth and T. Lange, "Feature selection in clustering problems," *Advances in Neural Information Processing Systems*, vol. 16, pp. 473–480, 2004.

[5] Ma. Dash, K. Choi, P. Scheuermann, and H. Liu, "Feature selection for clustering—a filter solution," *2002 IEEE International Conference on Data Mining, 2002. Proceedings*, 2002, pp. 115-122, doi: 10.1109/ICDM.2002.1183893.

[6] S. Visalakshi and V. Radha, "A literature review of feature selection techniques and applications: Review of feature selection in data mining," *IEEE International Conference on Computational Intelligence and Computing Research, 2014*, pp. 1-6, 2014, doi: 10.1109/ICCIC.2014.7238499.
[7] J. Tang, S. Alelyani, and H. Liu, "Feature selection for classification: a review," *Data classification: Algorithms and applications*, CRC Press, pp. 37–64, 2014.
[8] J. R. Vergara and P. A. Estévez, "A review of feature selection methods based on mutual information," *Neural Computing and Application*, vol. 24, no. 1, pp. 175–186, 2014, doi: 10.1007/s00521-013-1368-0.
[9] J. Xu, Z. Yuming, C. Lin, and X. Baowen, "An unsupervised feature selection approach based on mutual information," *Journal of Computer Research and Development*, vol. 49, no. 2, pp. 372-382, 2012.
[10] X. He, D. Cai, and P. Niyogi, "Laplacian score for feature selection," *Advances in Neural Information Processing Systems*, 2005, pp. 507-514.
[11] D. Cai, C. Zhang, and X. He, "Unsupervised feature selection for multi-cluster data," *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 333-342, 2010, doi: 10.1145/1835804.1835848.
[12] Y. Wang, "Unsupervised Representative Feature Selection Algorithm Based on Information Entropy and Relevance Analysis," in *IEEE Access*, vol. 6, pp. 45317-45324, 2018, doi: 10.1109/ACCESS.2018.2863752.
[13] L. Shi, L. Du, and Y. Shen, "Robust Spectral Learning for Unsupervised Feature Selection," *IEEE International Conference on Data Mining*, 2014, pp. 977-982, doi: 10.1109/ICDM.2014.58.
[14] F. Nie, W. Zhu, and X. Li, "Unsupervised feature selection with structured graph optimization," *Proceedings of the 30th Conference on Artificial Intelligence*, 2016, pp. 1302–1308, doi: 10.5555/3015812.3016004 .
[15] S. S.-Fernández, J. A. C.-Ochoa, and J. Fco. M.-Trinidad, "A review of unsupervised feature selection methods," *Artificial Intelligence Review*, vol. 53, pp. 907–948, 2020, doi: 10.1007/s10462-019-09682-y.
[16] S. Alelyani, J. Tang, and H. Liu, "Feature selection for clustering: a review," *Data Cluster Algorithms Applications*, pp.29-60, 2018.
[17] M. Dash and Y. Ong, "RELIEF-C: efficient feature selection for clustering over noisy data," *2011 IEEE 23rd International Conference on Tools with Artificial Intelligence*, 2011, pp. 869-872, doi: 10.1109/ICTAI.2011.135.
[18] V. M. Rao and V. N. Sastry, "Unsupervised feature ranking based on representation entropy," *2012 1st International Conference on Recent Advances in Information Technology (RAIT)*, 2012, pp. 421-425, doi: 10.1109/RAIT.2012.6194631.
[19] J. G. Dy and C. E. Brodley, "Feature selection for unsupervised learning," *Journal of Machine Learning Research*, vol. 5, pp. 845–889, 2004.
[20] P. Y. Lee, W. P. Loh, and J. F. Chin, "Feature selection in multimedia: the state-of-the-art review," *Image Vision Computing*, vol. 67, pp. 29–42, 2017, doi: 10.1016/j.imavis.2017.09.004.
[21] R. Varshavsky, A. Gottlieb, M. Linial, and D. Horn, "Novel unsupervised feature filtering of biological data," *Bioinformatics*, vol.22, no. 14, pp. e507–e513, 2006, doi: 10.1093/bioinformatics/btl214.
[22] M.i Banerjee and N. R. Pal, "Feature selection with SVD entropy: some modification and extension," *Information Sciences*, vol. 264, pp. 118–134, 2014, doi: 10.1016/j.ins.2013.12.029.
[23] S. S.-Fernández, J. F. Martínez-Trinidad, and J. A. Carrasco-Ochoa, "A new unsupervised spectral feature selection method for mixed data: a filter approach," *Pattern Recognition*, vol. 72, pp. 314–326, 2017, doi: 10.1016/j.patcog.2017.07.020.
[24] X. Zhu, Y. Wang, Y. Li, Y. Tan, G. Wang, and Q. Song, "A new unsupervised feature selection algorithm using similarity-based feature clustering," *Computational Intelligence*, vol. 35, no. 1, pp. 2-22, 2018, doi: 10.1111/coin.12192.
[25] H. Wang, Y. Zhang, J. Zhang, T. Li, and L. Peng, "A factor graph model for unsupervised feature selection," *Information Sciences*, vol. 480, pp. 144-159, 2019, doi: 10.1016/j.ins.2018.12.034.
[26] K. Torkkola, "Feature extraction by non-parametric mutual information maximization," *The Journal of Machine Learning Research*, vol. 3, pp. 1415-1438, 2003.
[27] J. Sainui and M. Sugiyama, "Direct approximation of quadratic mutual information and its application to dependence maximization clustering," *IEICE Transactions on Information and Systems*, vol. E96-D, no. 10, pp. 2282-2285, 2013, doi: 10.1587/transinf.E96.D.2282.

[28] J. Sainui and M. Sugiyama, "Unsupervised key frame selection using information theory and colour histogram difference," *International Journal of Business Intelligence and Data Mining*, vol. 16, no. 3, pp. 324-344, 2020, doi: 10.1504/ijbidm.2020.106137.

BIOGRAPHIES OF AUTHORS

Janya Sainui received her BS and MS in Computer Science from Prince of Songkla University, Hat Yai, Songkhla, Thailand, in 2005 and 2009, respectively. She is currently a Lecturer in Computer Science at Prince of Songkla University, Thailand. She has conducted research on topics such as bitmap indexing, video indexing and detection, clustering and unsupervised dimension reduction. She is interested in algorithm and application of machine learning, especially, applying machine learning techniques to real world applications such as image/video processing, medical imaging, bioinformatics, as well as information retrieval.

Chouvanee Srivisal is currently a Lecturer in Computer Science at Faculty of Science Prince of Songkla University, Thailand. She received her BS in Computer Science from Prince of Songkla University and MS in Information Technology from King Mongkut's Institute of Technology Ladkrabang, Thailand in 1997 and 2002, respectively. She has interested in algorithm and application of machine learning that for conducting research on topics such as image processing and information retrieval.