Computing and Informatics, Vol. 38, 2019, 223–239, doi: 10.31577/cai_2019_1_223

STUDY ON UNSUPERVISED FEATURE SELECTION METHOD BASED ON EXTENDED ENTROPY

Zhanquan Sun

Engineering Research Center of Optical Instrument and System
Ministry of Education, Shanghai Key Lab of Modern Optical System
University of Shanghai for Science and Technology, Shanghai, 200093, China
e-mail: sunzhq@sdas.org

Feng Li

Department of History, College of Liberal Arts, Shanghai University
Shanghai, 200436, China
e-mail: namelf@126.com

Huifen Huang

Shandong Yingcai University
Shandong, China
e-mail: shouyu1976@163.com

Abstract. Feature selection techniques are designed to find the relevant feature subset of the original features that can facilitate clustering, classification and retrieval. It is an important research topic in pattern recognition and machine learning. Feature selection is mainly partitioned into two classes, i.e. supervised and unsupervised methods. Currently research mostly concentrates on supervised ones. Few efficient unsupervised feature selection methods have been developed because no label information is available. On the other hand, it is difficult to evaluate the selected features. An unsupervised feature selection method based on extended entropy is proposed here. The information loss based on extended entropy is used to measure the correlation between features. The method assures that the selected features have both big individual information and little redundancy information with

the selected features. At last, the efficiency of the proposed method is illustrated with some practical datasets.

Keywords: Unsupervised feature selection, extended entropy, information loss, correlation value

1 INTRODUCTION

In recent years, data has become increasingly larger in number of features in many applications such as genome projects, text categorization, image retrieval and customer relationship management, etc. [1, 2]. It may cause serious problems to many machine learning algorithms with respect to scalability and learning performance. How to select the most informative variable combination is a crucial problem. Feature selection techniques are designed to find the relevant feature subset of the original features that can facilitate clustering, classification and retrieval [3, 4]. Feature selection is an important research issue in machine learning and pattern recognition. Lots of research work has been done on the topic. Based on whether the label information is available, feature selection is mainly partitioned into two types, i.e. supervised and unsupervised methods [5]. The former method is based on labeled samples for classification problems. The latter method is mainly used to analyze unlabeled data for clustering problems. Many supervised feature selection methods have been proposed and applied to many application areas. Typical supervised feature selection methods include correlation coefficient method, information gain, logistical regression, regularized method, etc. [6, 7, 8, 9]. In general, supervised feature selection is better in performance than unsupervised methods. But in practice, data samples are usually unlabeled. How to improve the performance of unsupervised feature selection is still a difficult problem to be resolved.

Supervised feature selection methods usually evaluate the importance of a feature by the correlation value between features and class variable. However, in practice, it is expensive or impossible to label large-scale samples in many applications. Hence, it is great significance to develop unsupervised feature selection algorithms that take full use of the unlabeled samples to select the most informative features. Some unsupervised feature selection methods have been proposed, such as maximum variance method, Laplacian score method, clustering based method, etc. [10, 11, 12]. For dealing with multi cluster feature selection problem, spectral regression and sparse space learning based method was proposed [13]. Feature selection is the process of selecting the most informative feature combination. The raw dataset contains many features that are either redundant or irrelevant. They can be removed without incurring much loss of information. Correlation metric is used to measure the relationship between features. Feature selection results based on different correlation metrics are different. Many kinds of correlation metrics have been proposed, such as Pearson correlation coefficient, mutual information and

so on. Mutual information can measure arbitrary statistical dependences between variables [14]. But the computational cost of mutual information between continuous variables and mutual information between discrete and continuous variables is expensive. Information bottleneck theory based information loss is an efficient correlation metric [15, 16]. It has been applied to many complicated clustering problems. But the information loss based on probability cannot process continuous variables. In this paper, information bottleneck theory based information loss is adopted to measure the correlation between features. For improving the general adapt capability, extended entropy is proposed and information loss is calculated based on it. The proposed feature selection method takes both the feature's individual entropy and the redundancy information with the selected features into consideration. It assures that the selected feature combination has the maximum information. For determining the number of selected features, an objective rule is proposed. The method combines the change ratio and the gradient ratio of information increase. At last, the efficiency of the proposed method is illustrated with some practical dataset.

The rest of this paper is organized as follows. Section 2 presents the definition of extended entropy. Section 3 introduces information bottleneck theory and information loss. Section 4 proposes the calculation method of information loss based on extended entropy. Feature selection procedure based on extended entropy is presented in Section 5. Some practical datasets are analyzed with the proposed method in Section 6. Concluding remarks are described in Section 7.

2 EXTENDED ENTROPY

Entropy is a way to measure the amount of information in a signal based on probabilities. Classic entropy is based on probability. Data has no statistical characteristics in many applications. A novel entropy definition, i.e. extended entropy, is proposed here. Extended entropy is not based on probability but on ratio. The definition is as follows.

2.1 Shannon Entropy

Let feature variables be denoted by vector $X = (X_1, X_2, \ldots, X_m)^T$, where $X_i = (x_{ij})$, $i = 1, 2, \ldots, m$, $j = 1, 2, \ldots, q$, denotes the i^{th} feature variable with q difference values, and class variable be denoted by Y, $Y = (y_i)$, $i = 1, 2, \ldots, k$. It means that all features are projected to k different classes. $p(X_i)$ denotes the probability distribution of feature variableX_i, p_Y denotes the probability distribution of class variable Y, and $p(X_i, Y)$ denotes the joint probability distribution of X_i and Y. All probability distributions are calculated according to sample statistics. The Shannon entropy H of feature variable X_i can be described as

$$H(X_i) = -\sum_{j=1}^{q} p(x_{ij}) \operatorname{logp}(x_{ij}). \tag{1}$$

Shannon entropy of class variable Y can be described as

$$H(Y) = -\sum_{i=1}^{k} p(y_i) \operatorname{logp}(y_i). \tag{2}$$

Joint entropy between feature variables and class variable is

$$H(X_i, Y) = -\sum_{j=1}^{q} \sum_{l=1}^{k} p(x_{ij}, y_l) \operatorname{logp}(x_{ij}, y_l) \tag{3}$$

where X_i can be substituted by subset of feature vector S, i.e., the joint entropy can be generalized to p variables.

2.2 Extended Entropy

Let N data vectors be denoted by y_i, $i = 1, 2, \ldots, N$, each vector has n positive number $y_{i1}, y_{i2}, \ldots, y_{in}$, $i = 1, 2, \ldots, N$ and the ratio between each positive number and the sum of all the positive number is

$$r(y_{ij}) = y_{ij}/(y_{i1} + y_{i2} + \ldots + y_{in}). \tag{4}$$

$r(y_{ij})$ is similar to the probability that satisfies $\sum_{j=1}^{n} r(y_{ij}) = 1$, and $r(y_{ij}) \geq 0$, $i = 1, 2, \ldots, n$. Extended entropy based on ratio is defined as

$$S(y_i) = -\sum_{j=1}^{n} r(y_{ij}) \ln r(y_{ij}). \tag{5}$$

3 INFORMATION BOTTLENECK THEORY

Information bottleneck (IB) theory is proposed to operate clustering problem. The theory is based on mutual information. The joint distribution of the object space X and the feature space Y is denoted by $p(x, y)$. According to the IB principle a clustering X that minimizes the information loss $I(X; \hat{X}) = I(X; Y) - I(\hat{X}; Y)$ is optimized. $I(X; \hat{X})$ is the mutual information between X and $\hat{X}$

$$I(X; \hat{X}) = \sum_{x;\hat{x}} p(x) p(\hat{x} \mid x) \log(p(\hat{x} \mid x)/(p(\hat{x}). \tag{6}$$

The IB principle is motivated from Shannon's rate-distortion theory which provides lower bounds on the number of classes. Given a random variable X and a distortion $d(x_1, x_2)$ measure, the symbols of X are represented with no more than R bits. The rate-distortion function is given

$$D(R) = \min_{p(\hat{x}|x) I(X;\hat{X}) \leq R} Ed(x, \hat{x}) \tag{7}$$

where $Ed(x,\hat{x}) = \sum_{x,\hat{x}} p(x)p(\hat{x} \mid x)d(x,\hat{x})$. The loss of the mutual information between X and Y caused by the clustering $\hat{X}$ can be viewed as the average of this distortion measure

$$\begin{aligned} d(x,\hat{x}) &= I(X;Y) - I(\hat{X};Y) \\ &= \sum_{x,\hat{x},y} p(x,\hat{x},y)\log(p(y \mid x))p(y) - \sum_{x,\hat{x},y} p(x,\hat{x},y)\log(p(y \mid \hat{x}))/p(y) \\ &= ED(p(x,\hat{x})||p(y,\hat{x})) \end{aligned} \tag{8}$$

where $D(f\|g) = E_f \log(f/g)$ is the Kullback-Lerbler divergence. The rate distortion function is

$$D(R) = \min_{p(\hat{x}|x)I(X;\hat{X})\leq R} (I(X;Y) - I(\hat{X};Y)) \tag{9}$$

which is exactly the minimization criterion proposed by the IB principle, i.e., finding a clustering that minimizes the loss of mutual information between the objects and the features. Let c_1 and c_2 be two clusters of symbols, the information loss due to the merging is

$$d(c_1,c_2) = I(c_1;Y) + I(c_2;Y) - I(c_1,c_2;Y), \tag{10}$$

information theory operation reveals

$$d(c_1,c_2) = \sum_{y,i=1,2} p(c_i)p(y \mid c_i)\log(p(y \mid c_i))/(p(y \mid c_1 \cup c_2)) \tag{11}$$

where $p(c_i) = |c_i| / |X|$, $|c_i|$ denotes the cardinality of c_i, $|X|$ denotes the cardinality of object space X, $p(c_1 \cup c_2) = |c_1 \cup c_2| / |X|$. It assumes that the two clusters are independent when the probability distribution is combined. The combined probability of the two clusters is

$$p(y \mid c_1 \cup c_2) = \sum_{i=1,2} |c_i| / (c_1 \cup c_2)p(y \mid c_i). \tag{12}$$

4 INFORMATION LOSS BASED ON EXTENDED ENTROPY

According to Equation (12), information loss based on probability can only process discrete variables. Therefore, the classic information loss definition is not suitable in many situations. Extended entropy can deal with any kind of positive dataset. We introduce extended entropy into information bottleneck theory. In the method, each element of the dataset y is taken as a different value probability of which is the ratio between each element's value and the sum of all the element's values. Let n samples and each sample include m features. Calculate the correlations between features according to the values in each sample. Each feature can be taken as an n dimension vector, i.e. $y_i = y_{i1}, y_{i2}, \ldots, y_{in}, i = 1, 2, \ldots, m$. Each sample is taken as a value of the feature variable. n samples means each feature has n values.

The extended probability of feature y_i is calculated according to the ratio between the feature value and the sum, i.e.

$$r(y_{ij}) = y_{ij}/(y_{i1} + y_{i2} + \cdots + y_{in}). \tag{13}$$

It can satisfy the conditions requirements, i.e. $\sum_{j=1}^{n} r(y_{ij}) = 1$ and $r(y_{ij}) \geq 0$, $j = 1, 2, \ldots, n$. The extended entropy based on extended probability is defined as

$$S(y_i) = -\sum_{j=1}^{n} r(y_{ij}) \ln r(y_{ij}). \tag{14}$$

The information loss due to the merging of two clusters is coherent to that of IB

$$d(c_1, c_2) = \sum_{i=1,2} \sum_{j=1}^{n} r(y_j \mid c_i) \log(r(y_j \mid c_i))/(r(y_j \mid c_1 \cup c_2)). \tag{15}$$

According to the calculation equation of information loss, after $p, q \in \{1, 2, \ldots, n\}$ being combined into a variable c, the extended probability of combine variable c can be denoted

$$r(y_{cj}) = \frac{|y_p|}{|y_p \cup y_q|} r(y_{pj}) + \frac{|y_p|}{|y_p \cup y_q|} r(y_{qj}). \tag{16}$$

5 FEATURE SELECTION BASED ON EXTENDED ENTROPY (FSBEE)

Unsupervised feature selection method FSBEE is as follows. A novel correlation definition is introduced. The correlation between feature variable X and feature variable Y is defined as

$$\rho(X, Y) = 1/d(X, Y). \tag{17}$$

The information loss value is inverse proportion to the correlation value. The features of Y are combined into one variable according to Equation (16) firstly when it is a feature set. Then the information loss $d(X, Y)$ is calculated according to Equation (15).

5.1 First Feature Selection

The first feature is selected according to the correlation value of each feature. The initial feature variable set is denoted by $X = \{X_1, X_2, \ldots, X_m\}$. U is used to denote unselected feature set and S is used to denote the selected feature set. At first, U is set the initial feature set and S is set null, i.e. $U = X$ and $S = \Phi$. The feature that has the biggest correlation value with the other feature subset is selected as the first one, i.e.

$$x_l = arg \max_{1 \leq i \leq m} \rho(X_i, (X \backslash X_i)). \tag{18}$$

The maximum correlation value means that the feature can represent the other features in maximum degree. The selected feature is added to selected feature set S and removed from unselected feature set, i.e. $S = \{X_l\}$ and $U = X \backslash X_l$.

5.2 Feature Selection Procedure

After the first feature has been selected, the other features are selected according to the following procedure. The k^{th} feature X_l is selected according to the increase of correlation value. The calculation of increase value is as follows.

$$X_l = \arg\max_{X_i \in U} \{\rho(X_i, (U \backslash X_i)) * d(X_i, S)\}, \tag{19}$$

$$f_k = \max_{X_i \in U} \{\rho(X_i, (U \backslash X_i)) * d(X_i, S)\}. \tag{20}$$

It means that the selected feature can represent the other unselected features in maximum degree. At the same time, the selected feature should provide the least redundancy information to the selected feature set S. The candidate feature should have the largest distance to the selected features. Then the selected feature X_l is added to the selected feature set S and removed from the unselected feature set U, i.e. $S = \{S, X_l\}$ and $U = U \backslash \{X_l\}$. Through iterating the above procedure, features are selected.

5.3 Determination of the Number of Selected Features

No objective rule is available to determine the number of selected features currently. It is often prescribed previously. In this paper, the number of selected features is determined according to the following rules. In the selection procedure, each step corresponds to an increase value of correlation. In general, the value will be in decreasing trend. The gradient ratio between the current step and the first step is calculated according to the following equation.

$$u = (f_{k-1} - f_k)/(f_1 - f_2). \tag{21}$$

The ratio between the increase value corresponding to the current step and that to the first step is denoted by

$$v = f_k/f_1. \tag{22}$$

Threshold values α for u and β for v are prescribed. When the values are less than prescribed threshold values, i.e. $u < \alpha$ and $v < \beta$, the feature selection procedure is stopped. The threshold values are prescribed according to a practical problem. Less threshold value corresponds to larger selected feature number. In general, the two threshold values are in the interval [0,1]. They are set to a less value when the analysis problem is complicated and to a bigger value to a simple problem. Pseudo-code of the feature selection procedure is summarized as Figure 1.

```
Main class
    Dataset load D. //Read data into matrix.
    Initialize variable vector U = X and S = ∅.
    Initialize α and β.
    For i←o to m
        Calculate correlation increase value f(i)
    End
    Obtain the maximum value f_max and its index k_max .
    Update S = {X_l} and U = X\X_l
      While (true)
          Calculate f(i), i ∈ U;
          Obtain f_max and its index k_max ;
          Update S = {S, X_l} and U = U\{X_l}.
          Calculating u and v
          If (u < α && v < β)
          Break;
      End
End mian class
```

Figure 1. Pseudo code of the feature selection procedure

From the above calculation procedure, the computation complexity of the feature selection procedure is about $O(knm^2)$, where k is the number of selected features, m is the number of total features, and n is the number of instances.

6 EXAMPLES

6.1 Data Source

The datasets are downloaded from the UCI machine learning website [17]. For proving the efficiency of the proposed unsupervised feature selection method, some classification datasets are analyzed. They are breast cancer clinic data, smart phone record, credit card record, mesothelioma data and image segmentation set. The basic information, i.e. number of features, number of instances, number of classes, of each dataset is listed in Table 1.

6.2 Feature Selection with the FSBEE

The label information is ignored during feature selection procedure. The label information is used to analyze the performance of feature selection method through comparing the classification results. For calculating the extended entropy, all features

Data Source	Number of Features	Number of Samples	Number of Classes
Breast cancer diagnostic data	30	569	2
Smart phone record	561	10 299	6
Credit card record	24	30 000	2
Mesothelioma data	34	324	2
Image segmentation data	19	2 210	7

Table 1. Dataset information

are transformed into positive values. In this example, all features are normalized into the interval $[0, 1]$. Then the features are selected according to the proposed method in Section 5. The selection results are as follows.

(1) Breast cancer diagnostic data. The data is provided by University of Wisconsin, Clinical Sciences Center. It is used for breast tumor diagnosis. The dataset includes 569 samples and each sample has 30 real-value features. They are categorized into two classes, i.e. benign or malignant. Firstly, each feature is normalized to the interval $[0, 1]$. Then, the extended probability of each sample can be calculated according to (13). The features are selected according to Section 5. For determining the number of selected features, the threshold value is set $\alpha = 0.1$ and $\beta = 0.1$. The increase of correlation value corresponding to each step is shown in Figure 2. At last, 14 features are selected.

(2) Smartphones dataset. The smartphones data is collected to recognize human activity [18]. 10 299 human activity records are collected and each record has 561 features. The human activity is categorized into 6 classes. Firstly, the features are normalized to the interval $[0, 1]$. Then features are selected according to the FSBEE method. The threshold values are set $\alpha = 0.01$ and $\beta = 0.01$. The correlation increase value of each step is shown in Figure 3. At last, 178 features are selected.

(3) Credit card record. The dataset is collected to identify the customer whether credible or not credible according to his/her personal payment record. There are 30 000 records in the dataset and each record has 24 attributes. Among the attributes, some are binary variables and some are continuous variables. For computing convenience, all features are normalized into interval $[0, 1]$ and taken as continuous variables. Then features are selected according to the FSBEE method. The threshold value is set $\alpha = 0.1$ and $\beta = 0.1$. The increase of correlation value of each step is shown in Figure 4. At last, 10 features are selected.

(4) Mesothelioma data. The dataset is collected by Dicle University from real patient reports. It is used to identify whether the patient's mesothelioma is benign or malignant. There are 324 samples in the dataset and each sample has 34 attributes. Firstly, all attributes are normalized to the interval $[0, 1]$. Then features are selected to according to the FSBEE method. The threshold value is

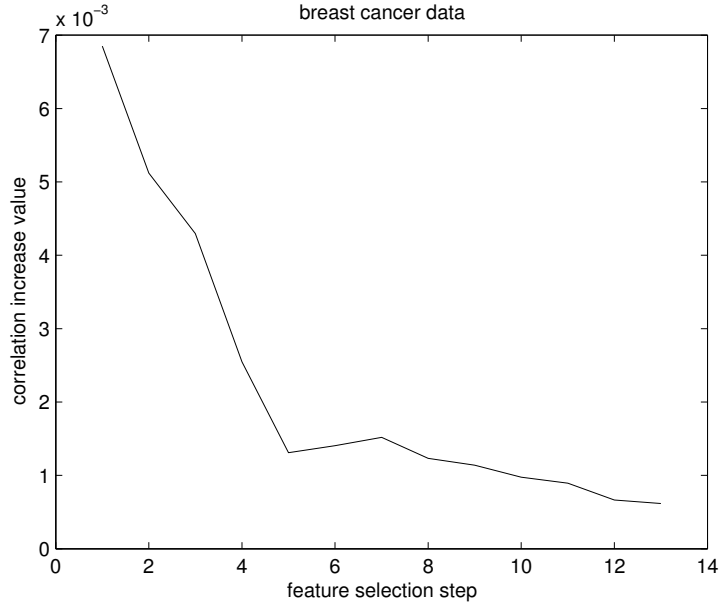


Figure 2. The increase in correlation value of breast cancer data

set $\alpha = 0.1$ and $\beta = 0.1$. The increase of correlation value of each step is shown in Figure 5. At last, 15 features are selected.

(5) Image segmentation. The instances were drawn randomly from a database of 7 outdoor images, i.e. brickface, sky, foliage, cement, window, path, grass. Each instance includes 19 features. The images were hand segmented to create a classification for every pixel. 210 images are taken as training set and 2 000 images are taken as test set. Firstly, all attributes are normalized to the interval $[0, 1]$. Then features are selected according to the FSBEE method. The threshold value is set $\alpha = 0.1$ and $\beta = 0.1$. The correlation increase value of each step is shown in Figure 6. At last, 12 features are selected.

6.3 Result Comparison

For comparison, information gain (IG), correlation coefficient, logistic regression etc. supervised feature selection methods are used to select the features. libSVM [19] is used to classify the dataset according to the selected features obtained with different feature selection methods. The number of selected features is the same as that of FSBEE model. The classification results are listed in Table 2. The computation time of the feature selection corresponding to each dataset are listed in Table 3. K-mean based feature selection, variance based feature selection and mutual infor-

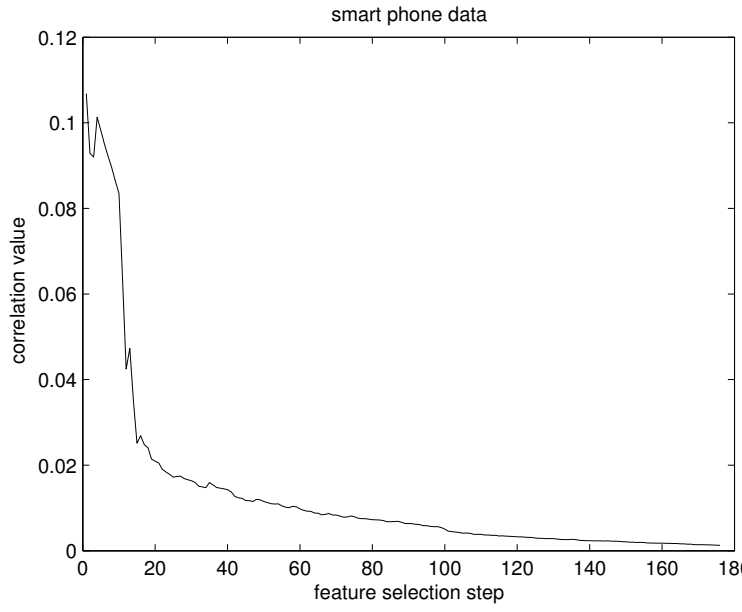


Figure 3. The increase in correlation value of breast cancer data

mation based feature selection (FSBMI) etc. unsupervised methods are used to select the features. The number of selected features is the same as that of FSBEE. The classification results based on different unsupervised feature selection method are listed as in Table 4. The computation time of the unsupervised feature selection corresponding to each dataset are listed in Table 5. From above analysis results we can find that the proposed FSBEE is an efficient unsupervised feature selection method. Through comparing with supervised feature selection methods, the classification accuracy of the proposed FSBEE method is close to that of the supervised method and even better than some supervised methods. Through comparing with other unsupervised methods, the classification accuracy of the proposed method is better. The proposed method is easy to be operated. The computation complexity will increase with the number of feature numbers and the number of selected feature numbers. But it is more efficient than the unsupervised feature selection method FSBMI.

7 CONCLUSIONS

Feature selection for unlabeled samples is a very important task for many pattern recognition problems. For improving the efficiency and performance of unsupervised feature selection, the paper develops a novel unsupervised feature selection method.

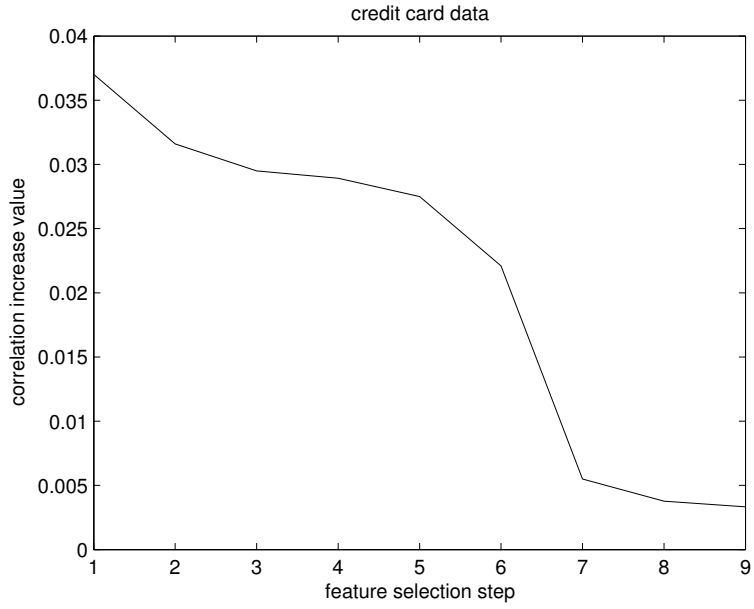


Figure 4. The increase in correlation value of breast cancer data

Data Source	Number of Selected Features	Accuracy			
		FSBEE	IG	Coefficient	Logistic Regression
Breast cancer diagnostic data	14	97.69	98.67	94.05	98.67
Smart phone record	178	89.47	91.32	91.06	90.75
Credit card record	10	77.46	78.68	78.11	77.43
Mesothelioma data	15	86.36	92.68	95.6	69.99
Image segmentation data	12	86.76	87.33	85.21	83.75

Table 2. Classification result comparison with supervised feature selection methods

Data Source	Number of Selected Features	Computation Time			
		FSBEE	IG	Coefficient	Logistic Regression
Breast cancer diagnostic data	14	0.156	0.016	0.015	1.228
Smart phone record	178	7455.375	4.563	4.422	6723.44
Credit card record	10	10.75	8.141	8.125	87.21
Mesothelioma data	15	0.406	0.313	0.297	4.563
Image segmentation data	12	0.047	0.012	0.01	0.53

Table 3. Computation time comparison with supervised feature selection methods

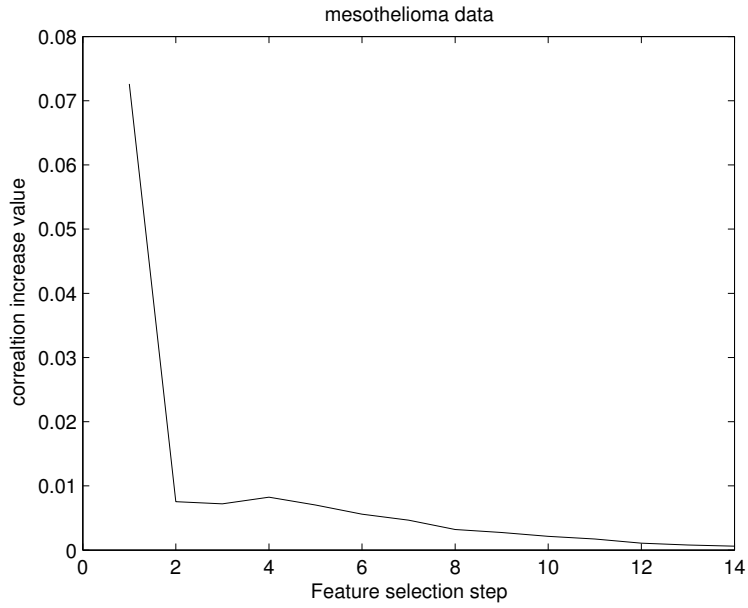


Figure 5. The increase in correlation value of breast cancer data

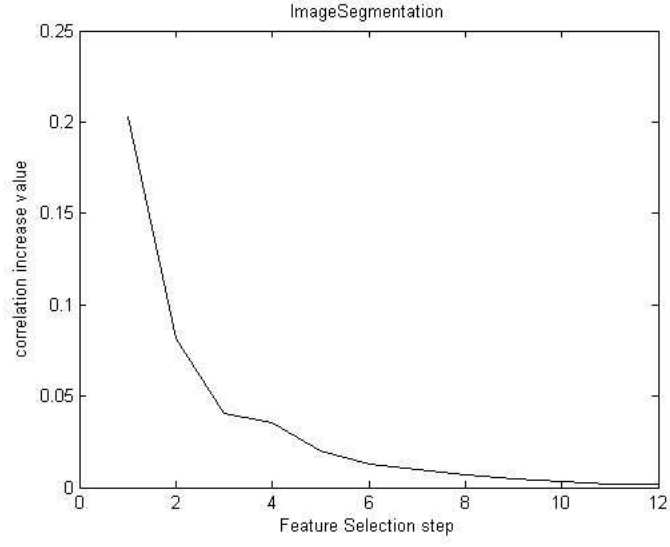


Figure 6. The increase of correlation value of image segmentation data

Data Source	Number of Selected Features	Accuracy			
		FSBEE	k-Mean Clustering Based Feature Selection	Variance Based Feature Selection	FSBMI
Breast cancer diagnostic data	14	97.69	96.84	94.39	100
Smart phone record	178	89.47	88.47	85.48	88.83
Credit card record	10	77.46	76.02	73.44	78.13
Mesothelioma data	15	86.36	89.61	84.65	76.95
Image segmentation data	12	87.76	80.21	78.43	81.98

Table 4. Classification result comparison with commonly unsupervised feature selection methods

Data Source	Number of Selected Features	ComputationTime			
		FSBEE	k-Mean Clustering Based Feature Selection	Variance Based Feature Selection	FSBMI
Breast cancer diagnostic data	14	0.156	0.24	0.031	0.781
Smart phone record	178	7455.375	154.33	4.359	9863.34
Credit card record	10	10.75	11.23	8.156	44.172
Mesothelioma data	15	0.406	0.43	0.297	0.797
Image segmentation data	12	0.047	0.049	0.125	0.14

Table 5. Computation time comparison with supervised feature selection methods

The method uses extended entropy to calculate the information loss between two features. It can measure the distance between features. During the feature selection procedure, the redundant information is considered. It assures that the selected features contain the maximum information. The advantages of the method can be summarized as three aspects. Firstly, extended entropy can simplify the calculation of information loss and improve computation speed markedly. Secondly, it can assure that the selected features contain the most information. Thirdly, it provides an objective rule to determine number of selected features. The experience results show that the proposed unsupervised method is efficient. It can be applied to many types of application areas.

Acknowledgements

This work is partially supported by the National Youth Science Foundation (No. 610-04115), Shandong Science and Technology Development Plan (No. 2016GGC01061), National Science Foundation (No. 61472230), National Youth Science Foundation (No. 61402271), the Natural Science Foundation of Shandong Province (Grant No. ZR2015JL023 and Grant No. ZR2015FL025).

REFERENCES

[1] Yu, L.—Liu, H.: Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution. Twentieth International Conference on Machine Learning (ICML '03), AAAI Press, 2003, pp. 856–863.

[2] Dash, M.—Liu, M.: Dimensionality Reduction. In: Liu, L., Özsu, M. T. (Eds.): Encyclopedia of Database Systems. Springer, 2009, pp. 843–846, doi: 10.1002/9780470050118.ecse112.

[3] Khalid, S.—Khalil, T.—Nasreen, S.: A Survey of Feature Selection and Feature Extraction Techniques in Machine Learning. 2014 Science and Information Conference (SAI), 2014, pp. 372–378, doi: 10.1109/sai.2014.6918213.

[4] Liu, X. M.—Tang, J. S.: Mass Classification in Mammograms Using Selected Geometry and Texture Features, and a New SVM-Based Feature Selection Method. IEEE Systems Journal, Vol. 8, 2014, No. 3, pp. 910–920, doi: 10.1109/jsyst.2013.2286539.

[5] Sarac, F.—Uslan, V.—Seker, H.—Bouridane, A.: Comparison of Unsupervised Feature Selection Methods for High-Dimensional Regression Problems in Prediction of Peptide Binding Affinity. 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2015, pp. 1–4, doi: 10.1109/embc.2015.7320291.

[6] Wang, D.—Nie, F. P.—Huang, H.: Feature Selection via Global Redundancy Minimization. IEEE Transactions on Knowledge and Data Engineering, Vol. 27, 2015, No. 10, pp. 2743–2755, doi: 10.1109/tkde.2015.2426703.

[7] Baccianella, S.—Esuli, A.—Sebastiani, F.: Feature Selection for Ordinal Text Classification. Neural Computation, Vol. 26, 2014, No. 3, pp. 557–591, doi: 10.1162/neco_a_00558.

[8] Sun, Z. Q.—Li, Z.: Data Intensive Parallel Feature Selection Method Study. 2014 International Joint Conference on Neural Networks (IJCNN), 2014, pp. 2256–2262, doi: 10.1109/ijcnn.2014.6889409.

[9] Sun, Z. Q.: Parallel Feature Selection Based on MapReduce. In: Wong, W. E., Zhu, T. (Eds.): Computer Engineering and Network. Springer, Cham, Lecture Notes in Electrical Engineering, Vol. 277, 2013, pp. 299–306, doi: 10.1007/978-3-319-01766-2_35.

[10] Hou, C. P.—Nie, F. P.—Li, X. L.—Yi, D. Y.—Wu, Y.: Joint Embedding Learning and Sparse Regression: A Framework for Unsupervised Feature Selec-

tion. IEEE Transactions on Cybernetics, Vol. 44, 2014, No. 6, pp. 793–804, doi: 10.1109/tcyb.2013.2272642.

[11] Aroquiaraj, I. L.—Thangavel, K.: Mammogram Image Feature Selection Using Unsupervised Tolerance Rough Set Relative Reduct Algorithm. 2013 International Conference on Pattern Recognition, Informatics and Mobile Engineering, 2013, pp. 479–484, doi: 10.1109/icprime.2013.6496718.

[12] Xu, J. L.—Zhou, Y. M.—Chen, L.—Xu, B. W.: An Unsupervised Feature Selection Approach Based on Mutual Information. Journal of Computer Research and Development, Vol. 49, 2012, No. 2, pp. 372–382, doi: 10.3724/sp.j.1087.2012.02250.

[13] Cai, D.—Zhang, C. Y.—He, X. F.: Unsupervised Feature Selection for Multi-Cluster Data. Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '10), 2010, pp. 333–342, doi: 10.1145/1835804.1835848.

[14] Witten, I. H.—Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques. 2nd ed. Morgan Kaufmann, Amsterdam, 2005.

[15] Battiti, R.: Using Mutual Information for Selecting Features in Supervised Neural Net Learning. IEEE Transactions on Neural Networks, Vol. 5, 1994, No. 4, pp. 537–550, doi: 10.1109/72.298224.

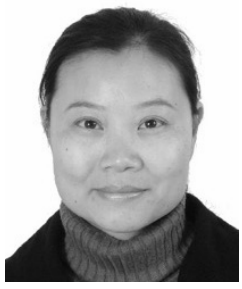
[16] Chiappino, S.—Marcenaro, L.—Regazzoni, C. S.: Information Bottleneck-Based Relevant Knowledge Representation in Large-Scale Video Surveillance Systems. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2014, pp. 4364–4368, doi: 10.1109/icassp.2014.6854426.

[17] UCI Machine Learning Repository. Center for Machine Learning and Intelligent Systems. http://archive.ics.uci.edu/ml/datasets.html.

[18] Anguita, D.—Ghio, A.—Oneto, L.—Parra, X.—Reyes-Ortiz, J. L.: A Public Domain Dataset for Human Activity Recognition Using Smartphones. 21th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, 2013, pp. 437–442.

[19] Fan, R.-E.—Chen, P.-H.—Lin, C.-J.: Working Set Selection Using Second Order Information for Training Support Vector Machines. Journal of Machine Learning Research, Vol. 6, 2005, pp. 1889–1918.

Zhanquan Sun, Ph.D., Associated Professor of University of Shanghai for Science and Technology. Major on big data, data mining and artificial intelligent. Has presided and attended 20 research projects and published about 60 academic papers.



Feng Li, Ph.D. candidate, Department of History, College of Liberal Arts, Shanghai University. Major on history data analysis. Attended about 10 research projects.

Huifeng Huang, Ph.D., Professor of Shandong Yingcai University. Major on image steganalysis, information security and data mining. Presided and attended 10 research projects and published about 30 papers.