Advances in Computer Science Research, volume 94
International Conference on Big Data, Electronics and Communication Engineering (BDECE 2019)

Unsupervised Feature Selection Algorithm Based on Information Gain

Zhong Li, Yang Jing*, Lijing Yao and Binbin Gan
School of Electrical and Electronic Engineering, North China Electric Power University, Baoding Hebei 071003, China
*Corresponding author

***Abstract*—Feature selection aims to select a smaller feature subset from the rate data which maintains the characteristics of the original data and has similar or better performance in data mining. traditional information theory often divides the relevance and redundancy of the features into consideration in unsupervised feature selection. This article proposes a supervised feature selection algorithm based on information gain analysis . this algorithm is to analyze the correlation between feature and original data and the redundancy between features and selected features based on the mutual information. The potential information gain of the feature is calculated for the feature sorting . At last, the feature is selected according to the gain penalty factor . The experimental results of multiple classifiers on multiple standard datasets show that the proposed algorithm achieves or better than the classification accuracy of the original data on the basis of effectively reducing the data dimension.**

Keywords—unsupervised; feature selection; mutual information; information gain

I. INTRODUCTION

With the development of big data and the widespread application of machine learning, people deal with massive amounts of high-dimensional data to obtain valuable information. Some of these features are redundant or unrelated, and they will affect the efficiency and performance of machine learning algorithms. It is an effective method to make the feature dimension of data lower by feature selection so that it can enhance the performance of machine learning algorithms.

A large number of researches have been made on feature selections. The results of the feature selections will be different duo to different methods. There are five methods of feature selection based on different methods of metrics: distance metric [1,2], correlation metric [3], consistency metric [4], information metric [5,6]and classification error rate metric [7]. Kira proposed a Relief feature selection algorithm using Euclidean distance [8]. The disadvantage is that only the relevant feature set can be selected but the redundant features cannot be removed. Qiu Like [9] proposed a feature selection algorithm based on redundancy analysis based on the correlation between correlation metrics and features and the correlation between features and categories. Various methods have been successfully applied in different fields, and the methods based on information metrics are widely used.

The feature selection problem can be divided into two categories based on whether the category information of the data to be processed is known. (1) supervised feature selection [10]. feature selection is performed under the condition of known data category information; (2) unsupervised feature selection [11]. Due to the lack of category information, it is difficult to select potential category related features. In recent years, scholars have conducted a lot of research on unsupervised feature selection methods. Ding Xuemei [12] proposed an unsupervised feature selection algorithm using improved ReliefF, which uses the DBSCAN clustering algorithm to guide classification, and improves the sampling strategy and applies the improved cosine similarity to measure the correlation of features to remove redundancy Characteristics. The determination of the parameter *m* of the improved ReliefF in this method has an influence on the feature evaluation, and it does not give a clear selection strategy in this article. Dy [13] proposed a feature selection algorithm based on clustering metrics, which uses the canonical clustering separability and likelihood to evaluate clustering results for different feature subsets with high time complexity. Xu Junling [14] proposed an unsupervised feature selection method based on mutual information (UFS-MI), and completed feature ranking based on unsupervised maximum correlation minimum redundancy criterion. However, it does not consider the amount of "new" information that the candidate feature brings to the selected feature set, and fails to give the dimension of the sorted feature selection.

This article proposes an unsupervised feature selection algorithm based on information gain due to the perspective of comprehensively considering the "new" information contained in the feature. The algorithm analyzes the correlation between features and original data and the redundancy between features based on mutual information technology, and calculates the information gain caused by the selection of selected features for feature selection.

II. INFORMATION THEORY

Entropy is often used to measure the uncertainty of variable information in information theory. The entropy of a discrete random variable $X=\{x_1,x_2,x_3,...x_n\}$ is denoted by $H(X)$, where $p(x_i)$ is the probability of the random variable X. $H(X)$ is defined as:

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log p(x_i) \quad (1)$$

For any two discrete random variables X and $Y=\{y_1,y_2,y_3,\ldots y_n\}$, the conditional entropy of the variable Y given X is defined as :

$$H(X \mid Y) = -\sum_{j=1}^{m} p(y_j) \sum_{i=1}^{n} p(x_i \mid y_j) \log p(x_i \mid y_j) \quad (2)$$

Where x_i and y_j represent the values of the random variables X and Y respectively. The value of $p(x_i \mid y_j)$ is the probability of the variable Y given X. If two random variables are independent of each other, the conditional entropy is equivalent to the information entropy. If not, the information entropy is greater than the conditional entropy.

Mutual information refers to the information shared by two variable, and is defined as:

$$I(X;Y) = H(X) - H(X \mid Y) = I(Y;X) \quad (3)$$

Mutual information can be expressed as the amount of information provided by variable X, which reduces the uncertainty of variable Y. And the value is equal to zero. if they are statistically independent.

III. Unsupervised Feature Selection Algorithm Based on Information Gain

In the supervised feature selection, we need to choose a satisfactory subset from the original data that can include almost all information and can effectively reduce redundancy. According to the information gain theory, we propose an unsupervised feature selection algorithm. Feature selection is performed by examining how much "new" information a feature brings.

According to the definition of mutual information mentioned above, this article uses mutual information to analyze the information existing in the original data set and the redundant information existing between the features. We use the information gain to analyze whether the feature is selected and complete the selection of the unsupervised feature subset. The relevant definitions are given below, and then the algorithm is described and illustrated.

A. Feature Relevance and Feature Redundancy

Suppose $S=\{f_1,f_2,f_3\ldots,f_n\}$ is the original data set of n-dimensional eigenvectors, where f_i is a certain dimensional feature vector. $U=\{u_1,u_2,u_3,\ldots,u_t\}$ is the pending feature set, where g_i is the pending feature. The definitions of "feature relevance " and " feature redundancy" in the feature selection process are given below.

Definition 1. Feature relevance. It expresses how much information the feature contains. The feature relevance of the feature is represented by the sum of the mutual information shared by all the features in the original data set in this article, so the feature relevance of the feature f_i is defined as：

$$\mathrm{Re}\,l(f_i) = \sum_{i=1}^{n} I(f_i; f_j) \quad (4)$$

Where $I(f_i; f_j)$ is the amount of information shared by the feature f_i and the feature f_j .

Definition 2. feature redundancy. It is expressed as redundant information between the candidate feature and the selected feature subset. The sum of the mutual information of all the features in the selected feature subset is used to represent the feature redundancy of the feature and the feature redundancy of the candidate feature g_i is defined as:

$$\mathrm{Re}\,d(g_i;U) = \sum_{1 \le k \le t}^{t} I(g_i;u_k) \quad (5)$$

Where $I(g_i;u_k)$ is the mutual information between the candidate feature g_i and the selected feature subset. As the candidate features are continuously added to the selected feature subset, more information is added to the selected feature subset, and the redundant information between the candidate feature and the selected feature subset is also increased, so the feature redundancy of the candidate feature is on the rise.

B. Feature Sorting and Selection

Based on the above definitions and analysis, this article considers that feature selection should choose those features that contain a large amount of potential information and have less redundant information shared with the selected feature subsets, which enables the feature subset to represent the original data to a greater extent. When selecting the first feature, we first calculate the feature relevance of each feature according to formula (4), and choose the feature with the highest feature relevance as the first important feature of feature selection. When selecting other features, we need to prioritize features that contain more information and less redundant information with selected features. Therefore, based on the theoretical analysis of information gain, this article gives the concept of “potential information gain” and applies it to the selection of other features.

Definition 3. Potential information gain. It represents the ratio of the potential information contained in the feature and the redundant information between the feature and the selected

subset, as well as the feature redundancy of the feature, so the potential information gain G is defined as:

$$G = \frac{\mathrm{Re}l(g_k) + \sum_{k<i\le n} \mathrm{Re}l(g_i)}{\mathrm{Re}d(g_k;U)} \quad (6)$$

We will select the follow-up other features one by one. We calculate the potential information gain of all the selected features according to formula (6), and select the feature corresponding to the maximum potential gain value.

Obviously, according to the theory of information gain, when the information that can be brought by the newly added feature is approximately equal to the redundant information when selecting features, it can be considered that the new feature cannot bring New information, thus the feature selection is terminated. Based on this idea, we give the definition of the gain penalty factor to determine whether features are added to the subset.

Definition 4 Gain penalty factor. It indicates the ratio of the maximum value of the potential information that may be brought by the candidate feature and the sum of redundancy information brought by the candidate feature and all the redundancy existing between the selected feature subset, so the Gain Penang factor C is defined as:

$$\mathrm{C} = \frac{\sum_{k\le i\le n} \mathrm{Re}l(g_i)}{\sum_{2\le j\le k}^{k} \mathrm{Re}d(g_j;U)} \quad (7)$$

Where $\sum_{2\le j\le k}^{k} \mathrm{Re}d(g_j;U)$ is the sum of the feature redundancy of all features in selected feature subset.

As described above, it is necessary to compare the feature relevance and feature redundancy of the feature f_i, to determine whether it is put into the subset. During the execution of the algorithm, the initial state feature subset is null, and the feature redundancy of the selected first feature is zero. With the dimension of the selected feature subset gradually increased, the feature redundancy of the candidate feature increases, and the gain penalty factor C decreases. Therefore, the gain penalty factor threshold is set to 1 as the criterion for evaluating whether a feature is put into the subset in the article. When the C value of the feature is 1, we consider that the amount of information contained in the candidate feature is substantially offset from the redundant information in the subset, and the "new" amount of information provided to the subset is "0".

Based on the above analysis, when selecting the feature subset, the penalty factor C is calculated by using formula (7) when judging whether the feature is added, and the algorithm termination condition is given by the threshold. The specific steps of feature selection are shown in Algorithm 1.

Algorithm 1: Feature selection algorithm based on information gain.

Input: Raw data set D=$\{f_1,f_2,f_3,\ldots,f_n\}$;
Output: Selected feature set S;
1. discretize $f_i \in$ D;
2. For each $f_i \in$ D
3. calculate the mutual information $I(f_i; f_j)$ of the features according to the formula(3)and mutual information matrix *I*;
4. calculate the feature relevance $Rel(f_i)$ of all features according to the formula (4),and $S(1) = \max(\mathrm{Re}l(f_i))$, then candidate feature subset is D=D-S;
5. End for
6. For 1<i⩽n
7. For each $f_i \in$ D
8. calculate *Red(f_i)* of the candidate features according to formula (5);
9. calculate *G(i)* of the candidate feature according to formula (6);
10. $S(i) = \max(G(i))$ and D=D-S;
11. calculate *C(i)* of the candidate feature according to formula (7);
12. End for
13. If C<1
14. Break;
15. End for
16. Output feature subset *S*

IV. Experiment Analysis

In the process of verifying the effectiveness of the proposed algorithm, we choose six data sets in the UCI machine learning data set, the parameters of which are shown in Table 1. In this article , the proposed algorithm is validated in three steps: 1) In the data preprocessing, different discretization methods have certain influence on feature selection, the Equal Width Split Box(EWSB) is used for discretization; 2) use the algorithm to select features for the discretized data set; 3) In the process of experimental analysis, K-nearest neighbor algorithm (KNN), decision tree algorithm and support vector machine classification algorithm (SVM) are used to verify the classification algorithm. We use the classification accuracy to verify the effectiveness of the algorithm in this article. We use the average value of the ten-fold cross-validation classification algorithm as the final result, so as to avoid the contingency of the algorithm.

TABLE I. DATA SET

No.	Dataset	No. Features	No. Instances	No. Classes
1	Iris	4	150	3
2	Liver disorders	6	345	2
3	Vehicle	18	846	4
4	Ionosphere	33	351	2
5	Sonar	60	208	2
6	Arrhythmia	262	452	16

We incrementally form the feature subsets of each data set one by one and input them into the classifier for classification experiments. As the dimension of the feature increases, the classification accuracy increases gradually. Figure I accurately reflects this trend. However, due to the increase of the feature dimension, the redundant information in the feature also increases gradually, which interferes with the classification, resulting in a decrease in accuracy.

Based on the above analysis, it is necessary to remove the unrelated features that interfere with the classification accuracy of the features, so that the selected subset can match with the complete set, even with the higher classification accuracy .

We use EWSB discretization method to preprocess the original data set, and then use the proposed algorithm to select features. We compare the dimensions and accuracy, and use three classification algorithms to verify that the algorithm effectively reduces the feature dimension while ensuring the classification accuracy of the selected subset. The results of various classifiers are shown in Table 2.

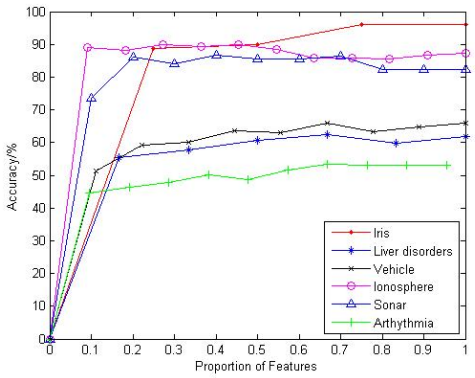

FIGURE I. 1-NN CLASSIFICATION ACCURACY RATE CORRESPONDING TO MULTIPLE DATA SETS

TABLE II. CLASSIFICATION ACCURACY OF VARIOUS CLASSIFIERS (EWSB)

Dataset	No. features		(KNN)Accuracy (%)		(DT) Accuracy (%)		(SVM) Accuracy (%)	
	Selected	***Complete***	***Selected***	***Complete***	***Selected***	***Complete***	***Selected***	***Complete***
Iris	3	4	**95.33**	95.33	**93.3**	93.3	**96.0**	96.0
Liver disorders	6	6	**61.70**	61.70	**62.9**	62.9	**70.1**	70.1
Vehicle	11	18	64.54	64.77	**70.4**	69.5	**74.9**	66.1
Ionosphere	23	33	**87.16**	86.88	86.9	88.6	**89.7**	88.3
Sonar	41	60	81.69	82.62	**73.6**	71.2	**88.5**	88.0
Arrhythmia	113	262	**53.09**	52.22	62.4	66.2	**62.4**	61.3

V. CONCLUSIONS

Feature selection is an important method for efficiently processing high-dimensional data. Aiming at the problem of how to comprehensively consider the new information and redundant information that can be brought about by features in the unsupervised feature selection process, we propose an unsupervised feature selection algorithm based on information gain. We sort the original data features based on the potential information gain, and choose the data features according to the gain penalty factor.

We choose the 10-fold cross-validation method to verify the effectiveness of the proposed algorithm, and we use some classifiers to prove that our algorithm is feasible. The selected subsets are compared with the complete set by the feature dimension and classification accuracy. The results show that while reducing the feature dimension, the classification accuracy of the selected subset is consistent with or even slightly improved, and it has a wider application prospect.

ACKNOWLEDGMENT

Project Supported by Science and Technology Project of State Grid Sichuan Electric Power Corporation(521908160001).

REFERENCES

[1] Pudil, P; Novovicova,J. Novel Methods for Subset Selection with Respect to Problem Knowledge. Intelligent Systems & Their Applications IEEE, 1998, 13(2):66-74.

[2] Robnik-Šikonja, M; Kononenko, I. Theoretical and Empirical Analysis of ReliefF and RReliefF. Machine Learning, 2003, 53(1-2):23-69.

[3] Wei,H L; Billings,S A. Feature Subset Selection and Ranking for Data Dimensionality Reduction. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2007, 29(1):162-6.

[4] Almuallim, H; Dietterich, T G. Learning Boolean Concepts in the Presence of Many Irrelevant Features. 1994.

[5] Sadia, S; Mohammad, S; Amin, A; Muhammad, A; Oksam, C. Simultaneous Feature Selection and Discretization Based on Mutual Information. Pattern Recognition,2019,91(91):

[6] Wang, H.J; Shi, Y.W; Wang, X. A Feature Selection Approach Based on Mutual Information Partitioning to Symbolic Data. Peak Data Scienc,2017,6(03):10-14

[7] Kohavi, R; John, G H. Wrappers for Feature Subset Selection. Artificial Intelligence, 1997, 97(1-2):273-324.

[8] Kira, K ; Rendell, L A . The Feature Selection Problem: Traditional Methods and a New Algorithm. // Tenth National Conference on Artificial Intelligence. AAAI Press, 1992.

[9] Qiu, L K; Guo, Z W; Liu, Q, Liu, Y J; Qiu, Z J, Feature Selection Algorithm Based on Redundancy Analysis, Journal of Beijing University of Posts and Telecommunications,2017,40(01):36-41

[10] Hien, D;Edwin, C; Patrick, S; Serge, W; Nicolas, B. Multi Criteria Series Arcfault Detection Based on Supervised Feature Selection. International Journal of Electrical Power and Energy Systems,2019,113.

[11] Renato, C. Unsupervised Feature Selection for Large Data Sets. Pattern Recognition Letters,2019

[12] Ding, X M; Wang, H J; Wang, Z G; Zhou, X Y, Unsupervised Feature Selection Method Based on Improved ReliefF, Computer Systems & Applications ,2018,27(03):149-155

[13] Dy, J G; Brodley, C E. Feature Selection for Unsupervised Learning. Journal of Machine Learning Research, 2004, 5(4):845-889.

[14] 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, 2012, 49(2):372-382.