IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON CONVERGENCE OF SENSOR NETWORKS, CLOUD COMPUTING, AND BIG DATA IN INDUSTRIAL INTERNET OF THING

Received December 27, 2017, accepted January 22, 2018, date of publication January 31, 2018, date of current version February 28, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2800287

An Industrial Internet of Things Feature Selection Method Based on Potential Entropy Evaluation Criteria

LONG ZHAO AND XIANGJUN DONG

School of Information, Qilu University of Technology (Shandong Academy of Sciences), Jinan 250000, China

Corresponding author: Long Zhao (zxcvbnm9515@163.com)

This work was supported in part by the Joint Special of Shandong Natural Science Foundation under Grant ZR2017LF020, in part by the National Natural Science Foundation of China under Grant 71271125 and Grant 61502260, and in part by the Natural Science Foundation of Shandong Province, China under Grant ZR2011FM028.

ABSTRACT In recent years, with the rapid development of industrial Internet of Things, the rapid growth of data has become a severe challenge and precious opportunity faced by many industries. The information society has entered the era of big data. Feature selection is frequently used to reduce the number of features in many applications of Internet of things, where data of high dimensionality are involved. To the best of our knowledge, a fewer researchers focus on the physical distribution of data and the anisotropy of the data characteristics. To this end, this paper introduces a novel feature selection approach based on potential entropy evaluation criteria (FMPE). The FMPE method considers the distribution of the data itself when measuring the importance of the feature. The data is mapped into a high-dimensional space which has better divisibility by extending data field to generalized multidimensional data field. Related experiments and analyses on UCI data sets and face data sets show that the FMPE algorithm can effectively eliminate the unimportant features or noise features to improve the performance of the classification algorithm. A high classification accuracy is achieved by the combination of the selected feature subset and a variety of classifiers and the FMPE algorithm is independent of the specific classifier.

INDEX TERMS Feature selection, potential entropy, data field, high classification accuracy.

I. INTRODUCTION

In the current Industrial Internet of things environment, the knowledge contained in it is excavated to guide actual production and specific application [1]. The importance of feature selection and learning is more prominent, which not only can effectively solve the ''Curse of dimensionality'', alleviate the current situation of ''information abundance, knowledge shortage'', but also better reduce complexity and understand data [2]. Duda *et al.* pointed out that the accuracy of classifier will decrease when the number of features exceeds the threshold [3]. The performance of classifier will decline sharply, especially when the correlation features are added. In the low dimensional space, the neural network with normal operation will fail as the dimension increasing [4]. Dimension reduction can effectively speed up the processing of the algorithm, avoid the bad effects of redundant features and noise features on pattern recognition [5]. With the increasing of the data dimension, in order to accelerate the speed of data processing, and avoid over fitting phenomenon, dimensionality reduction has become a hot research field of data mining [6]. In order to improve the speed of data processing and the pattern recognition accuracy dimension reduction is an important part of high-dimensional data mining.

The goal of dimension reduction is to find a low dimensional space in which data is organized into different clusters and easily separated. In addition, the low dimensional representation provides the possibility for data visualization and facilitates exploratory analysis of data [7]. In statistics, dimension reduction projects higher dimensional space to lower dimensional space, and get more accuracy of classification or regression. Set up a d dimensional data set R^d containing n samples, that is x_k $\{k = 1, 2 \ldots, n\}$, the goal of dimension reduction is to find a new projection space R^h, F: $R^d \rightarrow \mathrm{R^h}$, $x \rightarrow t = F(x)$ and t as the dimension reduction of x. The dimension of this space is $h(h < d)$, the point in R^h is t_k $\{k = 1, 2 \ldots, n\}$. Dimension reduction consists of two

strategies: feature selection and feature extraction [8]. Feature selection aims to select the most representative feature subset; the meanings of the selected features are not changed. The optimal feature subset is easy to understand. Feature extraction projects high-dimensional data to low dimensional space [9]. The new feature space extracted by feature extraction has more distinguishing ability than the original space, but the new feature space has no actual physical meaning. Relative to feature extraction, feature selection retains the origin form of features. This is one of the main reasons for feature selection is widely used. Since feature selection needs to measure the importance of all features and calculate the association between features, it does not apply to the dimension reduction of large number features data [10]. For this kind of data, feature selection is usually taken as the subsequent step of feature extraction to improve the generalization ability of classification model. Feature selection also is used to visualize and understand the data. The best subset of features can improve the efficiency and the accuracy of the classifier.

Dash and Liu point out that the classic feature selection consists of four basic steps: the generative process, the evaluation function, the stopping criterion, and the verification steps [11]. Among them, the generation process and the evaluation function are two main steps. The generative process is a search process for generating feature subsets. A variety of search strategies, including full search, heuristic search, and random search can be used to generate feature subsets. An evaluation function measures the identity of a feature subset and identifies different tags. The evaluation functions include five categories: distance, information, dependency, consistency, and classifier error rate. The feature selection algorithm using only the first four evaluation functions is called the filtered feature selection algorithm. Wrapper type feature selection algorithm uses classifier error rate as an evaluation function. In general, the Wrapper method can achieve better classification accuracy, but the filtering method is faster.

In filtering feature selection algorithms, the feature importance is obtained according to the intrinsic nature of data sets which is independent of subsequent algorithms. Numerous researchers have studied searching for a minimum subset of features which satisfy some goodness evaluation criterions. So far, various evaluation criteria have been studied to remove a number of less informative features. Such as, information entropy [12], correlation [13], rough set theory [14], clustering [15], constraint score [16], class separation [17], dependency measure [18] and consistency measure. For supervised classification processes, important features clearly contribute to the separation of features between classes.

Existing feature evaluation criteria have less research on the physical distribution of data and the relationship between data field features. Therefore, this paper proposes a new method to measure the feature importance based on potential entropy (FMPE). FMPE evaluates and analyzes the selection and application of the field function to measure the feature importance. It studies and improves the potential entropy selection strategy. Experiments and analyses in ten common data sets show that FMPE is independent of specific classification algorithm, and can select feature subset with high division which can effectively maintain or improve classification accuracy.

The rest of this paper is organized as follows. Fundamentals and the FMPE algorithm are presented in Section 2. Experiments and analysis of FMPE algorithm are showed in Section 3. Finally, conclusion and discussion are discussed in last Section.

II. THE FMPE ALGORITHM

The concept of generalized data field is introduced to FMPE algorithm, each original feature will be calculated according to potential value S_w within class and the potential value S_b between different classes. If the corresponding S_b is larger and S_w is smaller, this feature has a high class distinction and invariance within class.

Firstly, the data set is projected into the data space through the potential function, and the importance of the feature is calculated. Secondly, the more important steps are selection strategy of potential function, feature importance measurement based on potential entropy and feature subset search strategy.

A. SELECTION OF POTENTIAL FUNCTION

The estimation of potential value is related to the data set, the unit potential function and the influence factors. The anisotropy of data characteristics is mainly controlled by the influence of factor σ. Usually, the potential field distribution is unknown, but according to the potential function, the probability density function is a normalization constant. If the overall distribution of a given data set is known, the potential function estimation accuracy depends on the unit potential function and influence factors of the selection. When the influence factors are fixed, the same effect can be obtained for different potential functions. It is important to estimate the density of the potential function, that is, the center and magnitude of each data field.

Set $D = \{x_1, x_2, \ldots x_n\}$ is a data set, x_i is a data sample, and then the potential value of data point can be expressed as Equation 1.

$$\hat{\varphi}(x) = \sum_{i=1}^{n} m_i \times K(\frac{x - X_i}{\sigma}) \tag{1}$$

$K()$ is the potential function, m_i is the mass of xi, and σ is the factor. Where $\sum_{i=1}^{n} m_i = 1$. If each object has an equal mass m_i, a simplified formula for the potential function can be obtained as Equation 2.

$$\hat{\varphi}(x) = \frac{1}{n} \sum_{i=1}^{n} K(\frac{x - X_i}{\sigma}) \tag{2}$$

Suppose in the multidimensional data field, the influence factor sigma is different in different dimension, then the

potential function is defined as Equation 3.

$$\varphi(x) = \frac{1}{n}\sum_{i=1}^{n} m_i \times K\left(\frac{x - X_i}{H}\right) \tag{3}$$

where H is the $p \times p$ positive definite matrix which is related to the influence factor. In order to facilitate computation, the H is set a positive definite triangular matrix, and the simplified estimate of the potential function is defined as Equation 4.

$$\varphi(x) = \frac{1}{n}\sum_{i=1}^{n} m_i \times \left\{\prod_{j=1}^{p} K\left(\frac{x_j - X_{ij}}{\sigma_j}\right)\right\} \tag{4}$$

where σ_j is the influence factor of jth dimension. If the data is two-dimensional, then $j = 1, 2$.

B. FEATURE SELECTION BASED ON POTENTIAL ENTROPY

Given feature $f_k \in F', k = 1, 2, \ldots, p$ feature subset $F' \in F$. The potential value of data points in a data field is $\psi_1, \psi_2, \ldots \psi_n$, then the importance measure of the feature subset F' can be expressed as Equation 5.

$$Imp(f) = -\sum_{i=1}^{n} \frac{\psi_i}{Z} \cdot \log\left(\frac{\psi_i}{Z}\right) - \sum_{i=1}^{n}\left(1 - \frac{\psi_i}{Z}\right) \cdot \log\left(1 - \frac{\psi_i}{Z}\right) \tag{5}$$

where $\psi(x) = \frac{1}{n}\sum_{i=1}^{n}\left\{\prod_{F}' m_i \cdot k\left(\frac{x - X_i}{\sigma}\right)\right\}$, z is the sum of the potential values.

A feature potential entropy describes how much information a feature vector contains. The more information a feature contains, the smaller the entropy is. When the projection object is uniformly distributed, each object location potential value of approximately equal, and the corresponding feature importance tends to 0. On the other hand, the asymmetric distribution of projection object, entropy has a smaller potential value. When calculating the potential entropy of a data field, the minimum potential entropy corresponding to the optimal σ is defined as Equation 6.

$$Imp(f)_{opt} = \mathrm{imp}(f)_{\sigma = \sigma_{F'}} \tag{6}$$

In this paper, the nonparametric kernel density estimation method proposed by Hall P is used to calculate σ [19]. The standard deviation of the upper density of each direction is σ.

Another parameter m_i needed to calculate the potential value. FMPE uses the ratio of intra class distance and inter class distance as the quality of data points which is defined in Equation 7. Among them, the potential matrix is obtained by the potential function in the generalized data field. For supervised learning problems, the importance of the feature is related to the class potential value S_w and the inter class potential S_b.

$$m_i = \frac{S_{wi}^{\psi}}{S_{bi}^{\psi}} \tag{7}$$

where

$$S_{wi}^{\psi} = \frac{1}{n}\sum_{j=1}^{n}(\psi(X_i) - \psi_j(x))(\psi(X_i) - \psi_j(x))^T \tag{8}$$

The definition of spatial distribution matrix between classes S_b^{ψ} is as Equation 9.

$$S_{bi}^{\psi} = \frac{n_i}{n}(\psi(X_i) - \psi(X))(\psi(X_i) - \psi(X))^T \tag{9}$$

Total feature distribution matrix is defined as Equation 10.

$$S_t^{\psi} = S_w^{\psi} + S_b^{\psi} = \frac{1}{n}\sum_{i=1}^{n}(\psi(X_i) - \psi(X))(\psi(X_i) - \psi(X))^T \tag{10}$$

where $\psi(X) = \frac{1}{n}\sum_{i=1}^{n}\psi(X_i)$, $\psi_j(x) = \frac{1}{n_j}\sum_{i=1}^{n_j}\psi(X_i), j = 1, \ldots, c$. According to the above definition, S_w^{ψ} and S_b^{ψ} are non-negative matrix. For a given feature f normalization is also called deviation normalization. A linear transformation of the raw data maps the values to [0-1]. The translation function is defined as Equation 11.

$$f = \frac{f - min}{max - min} \tag{11}$$

max is the maximum of the sample data, and *min* is the minimum of the sample data. Mass vector $M = \{m_1, m_2, \ldots m_i, m_n\}$, where n is the number of features, and the value of m_i is the weight of the *ith* feature. The impact factor vector $\sigma = \{\sigma_1, \sigma_2 \ldots \sigma_i, \sigma_n\}$, σ_i is the factors affecting of the*ith* feature. The mean of each samples in all directions is the potential value of the sample that is defined as Equation 12.

$$\psi(X) = \frac{1}{n}\sum_{i-1}^{n}\psi(X_i) \tag{12}$$

where $\psi(X_i) = \prod_{F}' m_j \cdot k\left(\frac{x_{j-x_{i,j}}}{\sigma_j}\right)$, j is the *jth* sample of *ith* feature.

C. LAYER BY LAYER SEARCH OF IMPORTANT FEATURE SUBSETS

After getting out the feature importance, the best feature subset is obtained by using hierarchical clustering method [20]. The distance between the selected feature subset F′ and the label class C, $S_b(C; F')$ can be expressed as the sum of the distance between the selected features and the classed that is defined as Equation 13.

$$S_b(C; F') = \sum_{i=1}^{n} S(C_i; f) \tag{13}$$

Given the correlation between alternative features and selected features, that is:

$$S(f) = \sum_{f \in F'} \mathrm{CU}(f, s) \tag{14}$$

The update function within the class distance is defined as Equation 15.

$$S_W(F', s) = S_W(F') + S(f) \tag{15}$$

The size of the feature subset F' also needs to be considered. In general, the smaller the feature subset F' means the less selected features and more robust classifier. Based on

the above analysis, for each candidate feature, its evaluation function is defined as Equation 16.

$$J\left(F'\right) = \frac{S_b(C; F', s)}{|F'| + S_w(F', s)} \tag{16}$$

The $|F'|$ is the feature number of feature subset. The larger value of evaluation function $J(F')$ means the closer the correlation between the new feature and the class labels that the new feature subset is more helpful for classification. The evaluation function also takes into account the correlation between the selected feature subset and new candidate feature. If the correlation of candidate features and existing feature subset is too high, it means this feature is redundant and unnecessary. Instead, it shows that the new feature is an effective feature that can be incorporated into the feature subset. In this way, the final feature subset is guaranteed and the classification accuracy is improved and the feature subset is reduced.

A given feature subset of original feature set, properties of importance measure: if one is helpful to present the clustering structure features of $f \notin S$ was added. The distribution of data uncertainty will be reduced, then the minimum entropy generalized data field will become smaller, resulting in the importance of the new feature subset measure will be greater. On the contrary, if adding a confusing feature, the distribution of data uncertainty will increase, the minimum entropy corresponding to the generalized data field will increase, and the importance of the new feature will be reduced, that is $J(F' + f) < J(F')$. Obviously, an important feature for a subset F', delete any one of them could lead to a decline $J(F')$ with anti-monotonicity. Through the feature importance, the optimal feature subset is determined.

D. THE ALGORITHM FRAMEWORK OF FMPE

According to the information entropy important feature is added to the candidate feature subset. $S_b(C; F')$ and $S_w(F', s)$ of each candidate feature are calculated , the feature with largest $J(f)$ is combined with F' to assemble a new feature subset. If the selected feature number reaches the threshold or $Imp(F')$ becomes large, the selection process is end. The framework of the FMPE is shown in Algorithm 1.

If the sample data set contains n samples with m features, then the time complexity of the potential entropy calculation is $O(n^2)$. The time complexity of distance calculation between the candidate feature f and the tag class C is O(n). In the algorithm, the time complexity of the measurement standard $J(f)$ is O(nm). Therefore, the time complexity of select or assemble a candidate feature is O(nm^2). Then the total time complexity of FMPE is $O(n^2 + \mathrm{nm}^2)$.

III. EXPERIMENTS AND ANALYSIS

In order to verify the effectiveness of the proposed algorithm, experiments and analyses are carried out on ten typical datasets. Distance descriptions are also performed on the iris and teach data sets. Firstly, in the iris data set the projection images of different dimensions show that the high points of spatial data field. Secondly, on the teach data set through the distribution of potential lines to justify the feature selection results. Finally, the effectiveness and independence of FMPE algorithm are verified.

Algorithm 1 Feature Selection Method Based on Potential Entropy (FMPE)

Input: Training examples $\{x_1, x_2, \ldots x_n\}$ and class labels $\{y_1, y_2, \ldots y_c\}$, parameter δ

Output: Feature subset F;

1) Set $F' = \Phi$, $S_b = 0$, $S_w = 0$, δ;
2) For each f in F, Calculate its $Imp(f)$;
3) Sort F by $Imp(f)$ *value*;
4) $f = argmin\ (Imp(f))$; $F = F - \{f\} F' = \{f\}$;
5) While $\left|F'\right| < \delta$ or $J(F' + f) > J(F')$ do
   $f = argmin\ (Imp(f))$; $F' = F' + \{f\}$; $F = F - \{f\}$;
   $S_w = S_w + S(f)$, $S_b = S_b + S(C, f)$;
6) Return the subset F';

A. FEATURE SUBSET SELECTION OF IRIS

Taking Iris data as an example, the feature importance, the measurement process and the evaluation criteria of the feature subset are described. The Iris data set contains three classes: setosa, versicolor, and virginica. Each class contains 50 samples and the total is 150 samples.

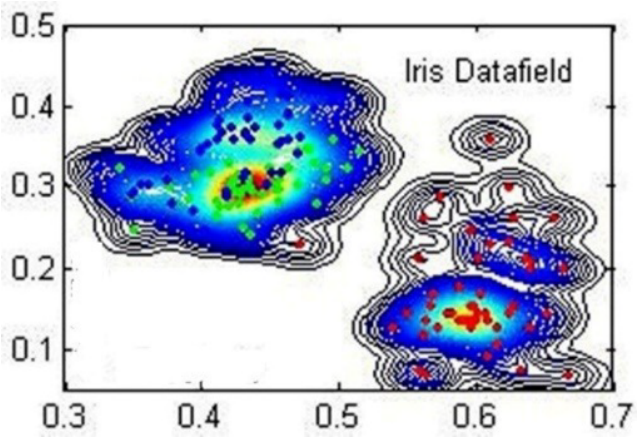

FIGURE 1. Two-dimensional data field graph of Iris.

The Iris data set is projected into the generalized data field by potential function. A two-dimensional map of the data field is shown in Fig.1. The three-dimensional map of the data field distribution is shown in Fig.2.

The Knn algorithm is used as the classifier on the original dataset and projected dataset. The result of classification is shown in Table 1.

According to the feature importance measure using generalized data field, the optimal feature subset of Iris is $\{X_3, X_4\}$. KNN classifier and PCA algorithm are used to extract the feature vector $\{pc_1, pc_2\}$, Misclassification sample number and rate of different subsets are shown in Table 2.

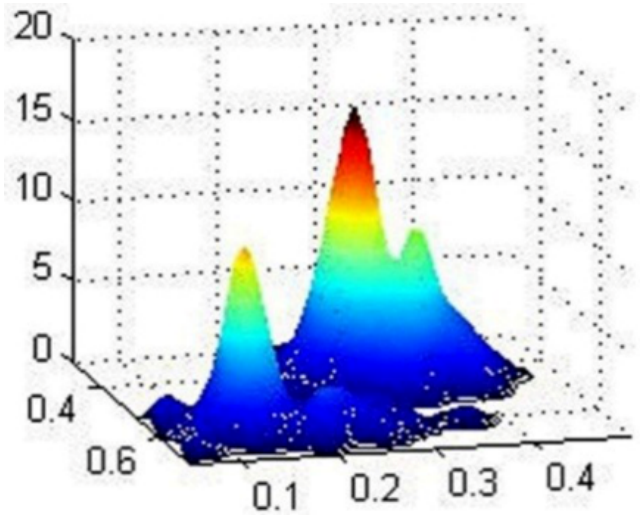

FIGURE 2. Three-dimensional data field graph of Iris.

TABLE 1. Error number and rate of Knn algorithm.

lable / data set	setosa	versicolor	virginica	Error rate
original	3	11	16	20.0%
projected	0	0	7	4.67%

TABLE 2. Classification accuracy of different feature subsets.

feature subset	$\{X_3, X_4\}$	$\{X_1, X_2, X_3, X_4\}$	$\{pc_1, pc_2\}$
Misclassification sample number	6	17	17
Misclassification rate	4%	11.33%	11.33%

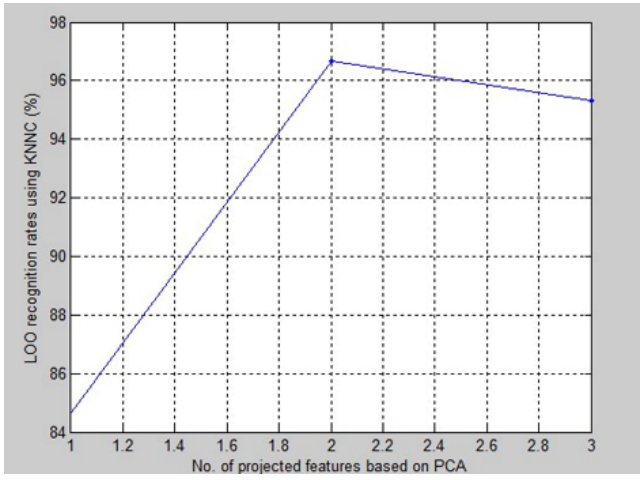

FIGURE 3. The classification accuracy of Iris using FMPE.

As the changes with the number of extracted features, the change of classification accuracy is shown in Fig 3.

Fig.3 is the classification accuracy varies with the dimensionality obtained by the KNN+FMPE. As can be seen, the FMPE method has the highest classification accuracy when the dimension number is two.

TABLE 3. Feature importance ordering of teach data sets.

Feature	X_5	X_2	X_3	X_1	X_4
importance	0.33	0.21	0.18	0.14	0.14

B. FEATURE SUBSET SELECTION OF TEACH

The teach data set contains five feature vectors. The metric of feature importance is calculated according to the potential entropy of the feature. Feature importance ordering of teach data sets is shown in Table 3.

Accordance with FMPE algorithm, the optimal feature subset is $\{X_2, X_3, X_5\}$. The histogram of the teach feature is shown in Fig.4.

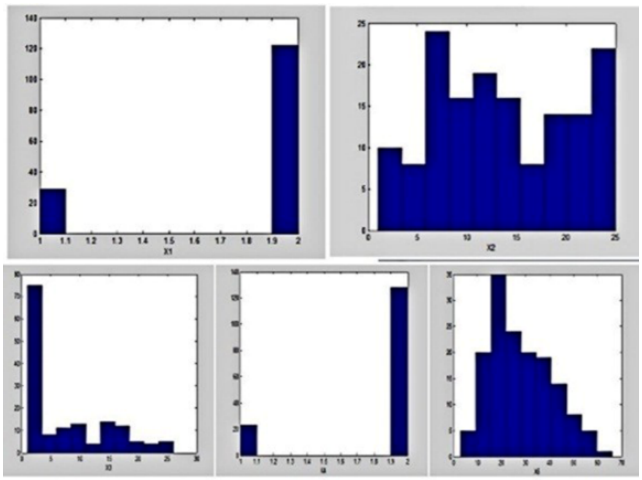

FIGURE 4. Histogram of teach features.

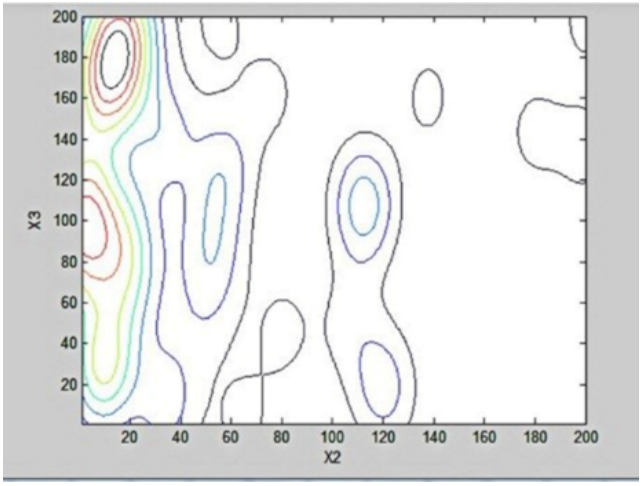

FIGURE 5. The equipotential lines of feature X_2 and X_4.

Through the analysis of Fig.4, the X_1 and X_4 features have two values, and the distinction between the different categories is small. X_2, X_3 and X_5 has high degree of distinction which is consistent with the above experimental results. The two-dimensional equipotential line distribution of feature X_2 and feature X_3 is shown in Fig.5.

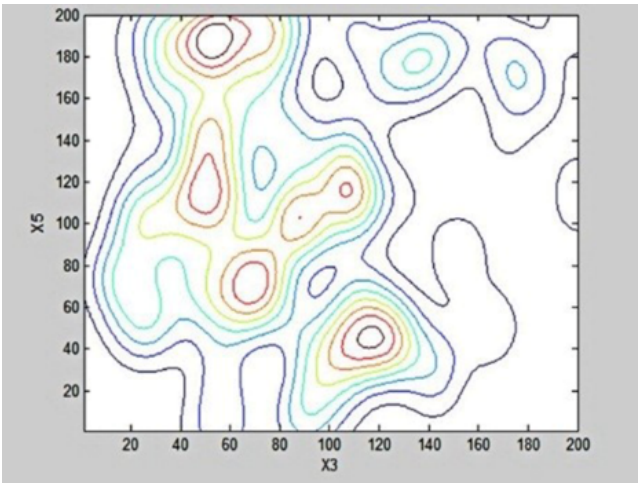

FIGURE 6. The equipotential lines of X_2, X_3 and X_5.

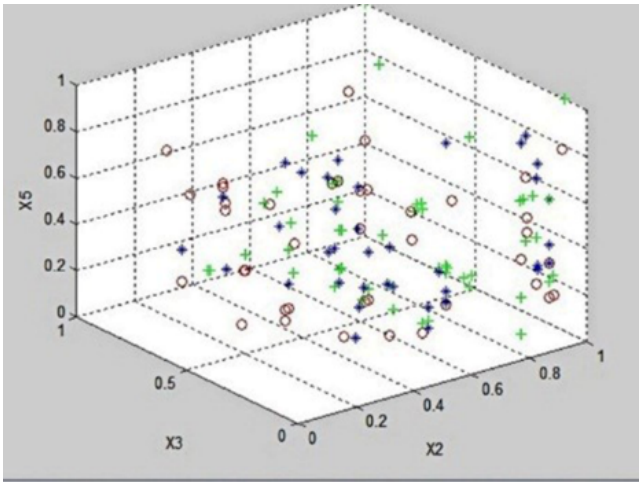

FIGURE 7. Scatter plots of X2, X3 and X5.

The two-dimensional equipotential line distribution of feature X_3 and feature X_5 is shown in Fig.6.

Fig 7 is a scatter diagram of feature X_2, X_3 and X_5. From the distribution of teach's 3D scatter diagram and its equipotential line, it can be seen from Fig.5, Fig.6 and Fig.7 that the teach data set embodies the local distribution feature, and automatically generates multiple field potential centers.

C. FMPE ALGORITHM INDEPENDENT EXPERIMENTS

In order to compare the classification performance of FMPE with other feature selection algorithms and identify its independent, ten public test data sets from UCI machine learning repository (http://archive.ics.uci.edu/ml) [21] were used as classification dataset. The basic properties of datasets are described in Table 4.

The mean filling strategy is used to fill the Incomplete data, that is to say, for missing or invalid data values, statistical interpolation is used to fill missing data. The FMPE algorithm is used to sort the feature importance of the data sets in Table 4. The feature importance ranking results are shown in Table 5.

TABLE 4. Dataset properties for experiments.

Data	Instances	Attributes	Classes
Cancer	198	32	3
Derm	366	33	6
Glass	214	9	6
Heart	270	13	2
Pro	997	20	3
Iris	150	4	3
Sonar	208	60	2
Teach	151	5	3
Wine	178	13	3
Vote	232	16	3

The order of feature importance is from small to large. The smallest feature importance measurement can be removed preferentially. The final feature is the largest feature importance measure, which allows priority to be added to feature subsets.

In order to verify the independence between the specific algorithm in the classification algorithm, this paper adopts three kinds of typical algorithm as the classifier, respectively is NBC (Naive Bayesian Classifier) [22], SRC (Sparse Representation based Classifier) [23] and SVM [24]. In order to make the algorithm more representative, the SRC algorithm adopts the SRC and SRCL respectively. The dimensionality and classification accuracy of each classifier are respectively given before and after dimensionality reduction. Firstly, the potential entropy of each feature is calculated and sorted, and then a feature subset is selected. 60% of the data set is used as training samples, and the rest is test samples. The accuracy of each algorithm is the average accuracy of ten times cross validation, and the arithmetic precision is retained at most one decimal places.

FMPE effectively removes the noise features that affect classification, and for NBC, SRC, SVM and SRCL classifiers, the classification accuracy is improved or almost unchanged after dimensionality reduction. The above experiments further confirm that the FMPE algorithm is independent of the specific classifier, which is only related to the distribution of the data itself, and also show that FMPE has a certain universality.

Through FMPE, the dimensionality of each data set is reduced to some extent, and the classification accuracy of classifier is improved. At the same time, due to the reduction of dimensionality, FMPE improves the generalization ability of classifier to a certain extent. We can find that the classification accuracy of the FMPE is higher than the prior dimension reduction in most datasets. It can be seen from Table 5 that the effect of dimensionality reduction is different in different data sets using FMPE algorithm. But after feature selection, the performance of classifier is basically improved. After reducing the dimension, the values of classification

TABLE 5. Classification accuracy of FMPE using different classifiers (%).

Data	Num	Num-FMPE	Nbc	Nbc-FMPE	Src	Src-FMPE	Svm	Svm-FMPE	Srcl	Srcl-FMPE
Cancer	32	16	66.7	64.7	73.3	**70.7**	72.7	72.7	60.6	**63.6**
Derm	33	29	96.7	96.2	97.3	97.3	98.9	97.3	23.5	**26.8**
Glass	9	7	50.5	47.7	68.2	67.3	75.7	**77.6**	64.4	**69.2**
Heart	13	6	59.5	**60.1**	50	**55.4**	62.8	60.8	36.5	**47.3**
Pro	20	17	65.3	65.1	90.7	90.4	69.1	**71.3**	87.8	87.4
Iris	4	2	96.0	96.0	94.7	94.7	97.3	96.0	92.0	**93.3**
Sonar	60	23	64.4	**67.3**	83.7	**89.4**	72.1	**76.0**	70.2	**85.6**
Teach	5	3	46.7	**48.0**	62.7	**64.0**	60.0	**64.0**	28.0	**30.7**
Vote	16	7	94.0	**95.7**	96.6	**97.4**	96.6	**97.4**	37.1	15.5
wine	13	6	94.0	**95.7**	80.9	**89.9**	47.2	**50.6**	95.5	93.3

TABLE 6. Feature importance ordering of wine data sets (%).

Feature	X_7	X_{10}	X_1	X_{12}	X_{13}	X_{11}	X_6	X_2	X_9	X_5	X_4	X_3	X_8
imp	15.9	12.7	9.1	9.0	8.4	7.9	7.8	6.7	5.6	4.9	4.4	4.2	4.0

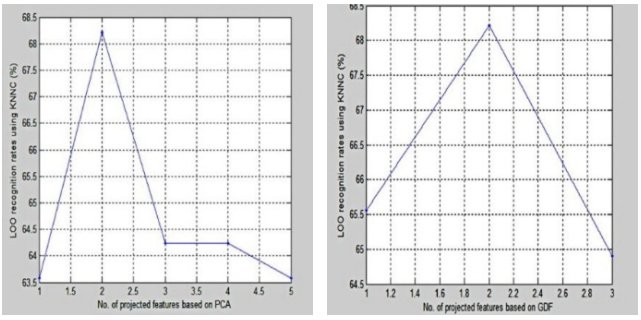

FIGURE 8. The classification accuracy changes of teach using PCA and FMPE.

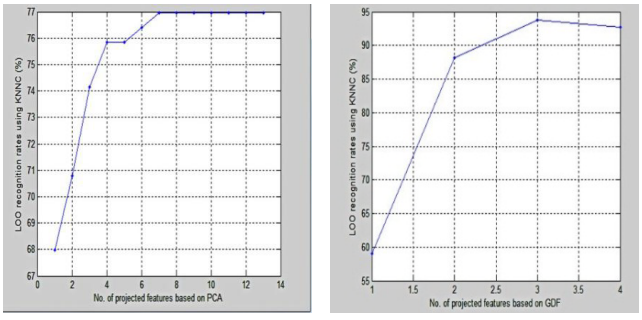

FIGURE 9. The classification accuracy changes of wine using PCA and FMPE.

accuracy are expressed in bold. Using the NBC classifier in the data set, five classification accuracy has been improved; the accuracy of the SCR classifier and the SVM classifier has been improved in six data sets; the accuracy of the SCRL classifier has been improved in seven data sets.

D. CLASSIFICATION EXPERIMENTS ON TEN UCI DATASETS

In order to further illustrate the effectiveness of the FMPE algorithm, the performance of FMPE algorithm and the classical dimensionality reduction algorithm PCA and LDA are compared and analyzed. Principal component analysis (PCA) is a multivariate statistical method, which is often used to reduce the dimensionality of multivariate signals [25]. The LDA algorithm can be used to determine the direction of projection. All samples are projected onto a coordinate axis, and the scatter between the classes of the projected samples in the projected feature space is the largest and the scatter within the class is minimum [26].

As the teach data set changes with the number of extracted features, the classification accuracy changes as shown in Fig.9. The classification accuracy of the KNN+PCA method varies with the dimension, and the right side of the image is the change of the classification accuracy after using the KNN+FMPE projection. As can be seen from Fig.8, both the PCA and the FMPE methods have the highest classification accuracy and the same classification accuracy when the dimension is two.

Wine data set was used for further experimental analysis. The 13 components of the data source wine are: Alcohol, Malic acid, Ash, Alkalinity of ash, Magnesium Tot, alphenols, Flavanoids, Nonflavanoid phenols, Proanthocyanins, Color intensity, Hue,OD280/OD315 of diluted wines, Proline. The importance metric ordering of each feature vector is shown in Table 6.

The data of each sample in the data source file is complete, and the change curve of the classification accuracy is shown in Fig.9.

The two-dimension reduction methods can effectively improve the classification accuracy, the PCA method in the 7 dimension is began to converge, the classification accuracy is 77%, the FMPE method in the 3 dimension is began to converge, the classification accuracy is 93.05%, regardless of the dimension selection or classification accuracy can be seen, the FMPE method is better than the PCA method.

In ten UCI data sets, respectively using FMPE and principal component analysis (PCA), linear discriminant analysis (LDA) for dimension reduction methods, combined

TABLE 7. Classification accuracies of different methods (%).

Data	KNN	Knn_LDA	Knn_PCA	Knn_FMPE
Cancer	76.26	68.12	65.70	**78.20**
Derm	**98.09**	97.00	95.80	98.08
Glass	69.16	73.36	73.72	**73.83**
Heart	46.13	41.41	41.80	**81.40**
Pro	91.07	84.70	88.20	**92.08**
Iris	96.67	96.70	96.00	**98.00**
Sonar	87.00	82.69	83.20	**89.53**
Teach	59.60	69.72	68.31	**70.20**
Vote	80.34	95.32	92.10	**97.80**

TABLE 8. Comparison of FMPE and other feature selection algorithms (%).

Data	KNN	SD	MI	RSFS	SFS	SFFS	FMPE
Cancer	76.26	74.24(10)	77.27(10)	75.76(11)	77.78(3)	77.78(3)	**78.20**(16)
Derm	98.09	84.7(10)	77.32(10)	92.35(12)	98.91(16)	96.72(9)	**98.08(29)**
Glass	69.16	68.69(6)	68.69(6)	72.90(5)	68.22(8)	71.03(7)	7**3.83**(7)
Heart	46.13	59.60(6)	59.60(6)	44.78(6)	57.24(8)	61.28(6)	**81.40**(6)
Pro	91.07	89.57(10)	89.57(10)	89.27(11)	**92.08**(17)	86.76(7)	**92.08**(17)
Iris	96.67	97.33(3)	97.33(3)	96.00(2)	96.67(4)	96.67(4)	**98.00**(2)
Sonar	86.54	86.54(10)	86.54(10)	87.98(21)	87.02(11)	**89.90**(16)	89.53(23)
Teach	59.60	59.60(5)	59.60(5)	59.60(3)	59.60(5)	59.60(5)	**70.20**(3)
Wine	80.34	82.58(7)	82.58(7)	82.58(6)	94.94(8)	96.07(7)	**97.80**(7)
Vote	93.53	96.98(6)	96.98(6)	96.98(5)	96.98(1)	**96.98**(1)	96.98(6)

with the KNN classifier for high-dimensional data sets. The classification accuracy was obtained after the classification, the classification accuracy as shown in Table 7.

The values in Table 7 are the classification accuracies corresponding to different dimensionality reduction strategies, and the maximum numerical accuracy of each of these four strategies is used for each set of data.

Among them, Knn_FMPE indicates that the FMPE is used to reduce the dimension first, and then the classification accuracy is obtained by using the KNN classifier. As can be seen from Table 7, KNN FMPE algorithm is better than other three algorithms in classification accuracy. In order to express this result intuitively, the classification result is described by column diagram as shown in Fig.10. It can be seen in Fig.10, the classification accuracy of FMPE in the 8 data sets are the highest, in the other two data sets on the classification accuracy and the highest classification accuracy shows that there is little difference between feature subset is obtained by the FMPE algorithm is more divisibility.

FMPE and the commonly used feature selection algorithms SD(Statistical Dependency) [27], MI(Manual Information) [28], RSFS(Random Subset Feature Selection) [29], (Sequential Floating Forward Selection) [30] are compared and tested. The comparison results are shown in Table 8.

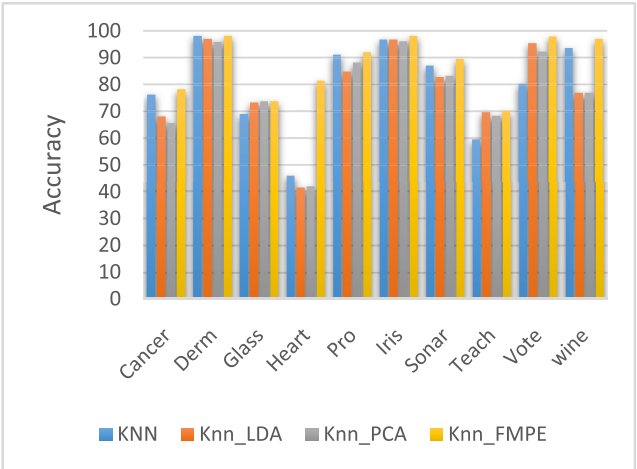

FIGURE 10. Classification accuracy contrast chart.

In Table 8, still using the UCI public data sets are related to the experimental analysis, the first column is the classification accuracy using the KNN algorithm in the original data set; behind each column indicate the use of different feature selection algorithm to obtain the optimal feature subset, the classification accuracy by KNN classification, the classification accuracy of two decimal places.

TABLE 9. The number of features selected by FMPE.

Dataset	Fnum	FMPE
Umist	644	38
Orl_original	10304	58
AR_database	19800	103

TABLE 10. Classification accuracy comparison of different algorithms (%).

Datasets / Method	Umist	Orl_original	AR_database
pca	99.5	98	92.1
sift	95.1	91.5	97.1
pso	95.1	90.5	94.3
cmim	94	93.2	92.1
AdaBoost	99.3	98.5	90.7
FMPE	99.7	98.7	98.8

The numerical value contained in the posterior bracket is the feature number of the feature subset. It can be seen, the classification accuracy on ten data sets of FMPE are more than the most feature selection strategy. FMPE can choose a subset of features high discrimination, but also should pay attention to that the average number of extracted features by FMPE algorithm is more than other algorithms.

E. CLASSIFICATION EXPERIMENTS ON THREE FACE DATASETS

Face data is a typical high-dimensional data, and how to find the feature points of face has been a hot research area of dimensionality reduction. FMPE algorithm uses the form of potential entropy to represent the features of human face data, and finds the points with great potential energy as the recognition feature for face recognition. In order to verify the effectiveness of the algorithm, this paper does some experiments and Analysis on three general face databases ORL, Umist and AR_database.

Table 9 shows the number of features selected by the original dimension Fnum and FMPE algorithm in three general face databases.

The accuracy of face recognition using 7 dimensionality reduction methods is shown in Table 10.

It can be seen, the classification accuracy of FMPE are more than the most feature selection strategy. FMPE can choose a subset of features with high discrimination.

IV. CONCLUSION

New technology changes, such as the end to end transmission of the Industrial Internet of things, have pushed mankind to an era of great information [31]. However, in the face of the electronic information of the multitude, they are at a loss, how to acquire the information we need is an urgent problem [32]. In feature selection, eliminating important features can lead to lower performance of the learning algorithm. For supervised classification, an efficient feature subset selection mechanism can improve the performance of classifier. The importance of features is an important basis for feature selection, according to the importance of the existing feature measurement algorithm on the features of physical distribution and spatial distribution of sample points to consider fewer problems, this paper puts forward a new measurement algorithm based on entropy FMPE features of potential importance. FMPE calculates the entropy of each feature by the intra class and inter class potential values. According to the hierarchical clustering algorithm, select those relatively important features, forming the best subset of features. Compared with FMPE and PCA, LDA algorithm, the overall performance of FMPE is better than the latter two algorithms. Combining FMPE with a variety of classifiers, the classifier can effectively improve or maintain the classification accuracy of the classifier on the basis of reducing the data dimensionality. Comparing the FMPE and the commonly used feature selection algorithms, the average classification accuracy of the feature subset selected by FMPE is the highest. Although the classification accuracy of FMPE is satisfactory, there is still room for improvement in terms of the data specificity and the selection of feature subsets [33]. However, the proposed FMPE and related experimental analysis also proved the feasibility of using the potential entropy to measure the importance of features, which laid the foundation for further research.

REFERENCES

[1] J. Li, L. Huang, Y. Zhou, S. He, and Z. Ming, "Computation partitioning for mobile cloud computing in a big data environment," *IEEE Trans. Ind. Informat.*, vol. 13, no. 4, pp. 2009–2018, Aug. 2017.
[2] J.-Q. Li, F. R. Yu, G. Deng, C. Luo, Z. Ming, and Q. Yan, "Industrial Internet: A survey on the enabling technologies, applications, and challenges," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 3, pp. 1504–1526, 3rd Quart., 2017.
[3] R. O. Duda, P. E. Hart, and D. G. Stork, "Pattern classification," in *En Broeck the Statistical Mechanics of Learning Rsity*, 2nd ed., 2000.
[4] V. Tangkaratt, H. Sasaki, and M. Sugiyama, "Direct estimation of the derivative of quadratic mutual information with application in supervised dimension reduction," *Biopolymers*, vol. 29, no. 8, pp. 2076–2122, 2017.
[5] K. Cui and T. T. Zhao, "Unsaturated dynamic constitutive model under cyclic loading," *Cluster Comput.*, vol. 20, no. 4, pp. 2869–2879, 2017.
[6] B. Xue, M. Zhang, W. N. Browne, and X. Yao, "A survey on evolutionary computation approaches to feature selection," *IEEE Trans. Evol. Comput.*, vol. 20, no. 4, pp. 606–626, Aug. 2016.
[7] Y. Sun, H. Qiang, X. Mei, and Y. Teng, "Modified repetitive learning control with unidirectional control input for uncertain nonlinear systems," *Neural Comput. Appl.*, pp. 1–10, 2017.
[8] M. A. Ambusaidi *et al.*, "Building an intrusion detection system using a filter-based feature selection algorithm," *IEEE Trans. Comput.*, vol. 65, no. 10, pp. 2986–2998, Oct. 2016.
[9] F. Zhang, P. P. K. Chan, B. Biggio, D. S. Yeung, and F. Roli, "Adversarial feature selection against evasion attacks," *IEEE Trans. Cybern.*, vol. 46, no. 3, pp. 766–777, Mar. 2016.
[10] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, Mar. 2003.
[11] M. Last, A. Kandel, and O. Maimon, "Information theoretic algorithm for feature selection," *Pattern Recognit. Lett.*, vol. 22, pp. 799–811, May 2001.
[12] T. F. Covões, E. R. Hruschka, L. N. de Castro, and Á. M. Santos "A cluster-based feature selection approach," in *Proc. 4th Int. Conf. Hybrid Artif. Intell. Syst.*, 2009, pp. 169–176.

[13] F. Jiang, Y. Sui, and L. Zhou, "A relative decision entropy-based feature selection approach," *Pattern Recognit.*, vol. 48, pp. 2151–2163, Jul. 2015.
[14] S. Shilu, K. Sheth, and E. Mehul, "Implementation of FAST clustering-based feature subset selection algorithm for high-dimensional data," in *Proceedings of International Conference on ICT for Sustainable Development*. Singapore: Springer, 2016.
[15] A. Alalga, K. Benabdeslem, and N. Taleb, "Soft-constrained Laplacian score for semi-supervised multi-label feature selection," *Knowl. Inf. Syst.*, vol. 47, no. 1, pp. 75–98, 2016.
[16] C. Silva, T. Bouwmans, and C. Frélicot, "Online weighted one-class ensemble for feature selection in background/foreground separation," in *Proc. Int. Conf. Pattern Recognit.*, Dec. 2016, pp. 2216–2221.
[17] M. S. Raza and U. Qamar, "An incremental dependency calculation technique for feature selection using rough sets," *Inf. Sci.*, vols. 343–344, pp. 41–65, May 2016.
[18] K. Shin and S. Miyazaki, "A fast and accurate feature selection algorithm based on binary consistency measure," *Comput. Intell.*, vol. 32, no. 4, pp. 646–667, 2016.
[19] P. Hall, S. Sheather, M. C. Jones, and J. S. Marron, "On optimal data-based bandwidth selection in kernel density estimation," *Biometrika*, vol. 78, no. 2, pp. 263–269, 1991.
[20] J. Hua, W. D. Tembe, and E. R. Dougherty, "Performance of feature-selection methods in the classification of high-dimension data," *Pattern Recognit.*, vol. 42, no. 3, pp. 409–424, 2009.
[21] A. Asuncion and D. J. Newman, "UCI machine learning repository," School Inf. Comput. Sci., Univ. California, Irvine, Irvine, CA, USA, 2007.
[22] A. Bender *et al.*, "Molecular similarity searching using atom environments, information-based feature selection, and a naive Bayesian classifier," *J. Chem. Inf. Comput. Sci.*, vol. 44, no. 1, pp. 170–178, 2004.
[23] L. Zhang *et al.*, "Kernel sparse representation-based classifier," *IEEE Trans. Signal Process.*, vol. 60, no. 4, pp. 1684–1695, Apr. 2012.
[24] V. N. Vapnik, *The Nature of Satistical Learning Theory*. New York, NY, USA: Springer-Verlag, 1999, pp. 988–999.
[25] I. T. Jolliffe, *Principal Component Analysis*, vol. 8. Berlin, Germany: Springer, 2010, pp. 41–64.
[26] D. L. Swets and J. Weng, "Using discriminant eigenfeatures for image retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 18, no. 8, pp. 831–836, Aug. 1996.
[27] B. Chandra and M. Gupta, "An efficient statistical feature selection approach for classification of gene expression data," *J. Biomed. Informat.*, vol. 44, no. 4, pp. 529–535, 2011.
[28] M. A. Sulaiman and J. Labadin, "Feature selection based on mutual information," in *Proc. Int. Conf. Asia IEEE*, Aug. 2015, pp. 1–6.
[29] J. M. Cadenas, M. C. Garrido, and R. Martínez, "Feature subset selection filter-wrapper based on low quality data," *Expert Syst. Appl.*, vol. 40, no. 16, pp. 6241–6252, 2013.
[30] A. Marcano-Cedeño, J. Quintanilla-Domínguez, M. G. Cortina-Januchs, and D. Andina, "Feature selection using sequential forward selection and classification applying artificial metaplasticity neural network," in *Proc. 36th Annu. Conf. IEEE Ind. Electron. Soc. (IECON)*, Nov. 2010, pp. 2845–2850.
[31] A. Yang *et al.*, "Optimum surface roughness prediction for titanium alloy by adopting response surface methodology," *Results Phys.*, vol. 7, pp. 1046–1050, 2017.
[32] K. Cui, W. Yang, and H. Gou, "Experimental research and finite element analysis on the dynamic characteristics of concrete steel bridges with multi-cracks," *J. Vibroeng.*, vol. 19, no. 6, p. 41984209, 2017.
[33] W. Wei *et al.*, "Gradient-driven parking navigation using a continuous information potential field based on wireless sensor network," *Inf. Sci.*, vol. 408, pp. 100–114, Oct. 2017.

LONG ZHAO received the M.S. degree in computer science and technology from Shandong Polytechnic University in 2009 and the Ph.D. degree from Wuhan University in 2016. He is currently a Lecturer with the School of Information, Qilu University of Technology (Shandong Academy of Sciences). His research interests include image processing, machine learning, and knowledge discovery.

XIANGJUN DONG received the M.E. degree in computer applications from Shandong Industrial University in 1999 and the Ph.D. degree in computer applications from the Beijing Institute of Technology in 2005. He is currently a Professor with the School of Information, Qilu University of Technology (Shandong Academy of Sciences), Jinan, China. His research interests include association rules, sequential pattern mining, and negative sequential pattern mining.