FEATURE SELECTION ALGORITHM BASED ON CONDITIONAL DYNAMIC MUTUAL INFORMATION

Wang Liping

Engineering Technology Research Center of Optoelectronic Technology Appliance, AnHui Province, TongLing, 244000, China

Submitted: Oct. 30, 2014 ***Accepted: Jan. 14, 2015*** ***Published: Mar 1, 2015***

Abstract- Aim at existing selection algorithm mutual information inaccurate valuation problem, a condition dynamic concept of mutual information. On this basis, the conditions proposed based on dynamic mutual information (CDMI) feature selection algorithm to overcome the traditional mutual information selection process dynamic correlation problem; conditions of dynamic mutual information throughout the selection process is dynamic valuation, those the samples can be identified after each selection features removed so that they no longer participate in conditional mutual information calculation process, accurate measurement sample. Accurate measurement sample on the degree of importance characteristics and at the same time ensure that the characteristics of information content. The experimental results verify the correctness and effectiveness of the algorithm.

Index terms*:* Dalgaard-Strulik model, energy, economic growth, time delay, limit cycle.

I. INTRODUCTION

As the Internet and databases and other information technology rapid development, many areas or sectors produced and accumulated a large amount of data. These are not only a large number data, and the data used to represent the characteristics (or properties) is very high number. Knowledge refers to data processed and refined a form of expression, it is for humans to understand and transform the objective world plays a vital role. Traditional knowledge acquisition method is generally intuitive way by manually or directly from the data model to build awareness, to get people to understand useful information. This manual acquisition method for small amounts of data is very useful in terms of efficiency is relatively high, but when faced with a large number, even when large amounts of data, its limitations it clearly exposed, so that can't meet the information needs of rapid development. This is called the "data explosion, lack of knowledge" phenomenon . Therefore, how to effectively deal with data, and data from the mass to find or find people available knowledge, is the cross in front of people's problems. This is also the main contents of information processing.

As a multidisciplinary field of research , knowledge discovery (also known as data mining) , machine learning and pattern recognition, and so is the intelligent information processing techniques reflect different methods , including data classification of these research areas is the main direction of specific topics or a . Although these studies have their different techniques and starting point for research purposes , but they all have the same or similar data processing , from existing or historical data for training to learn and dig out potentially useful model, which can be extracted knowledge or guidance to describe the user's behavior [1] .

Typically, large-scale data set contains a lot of irrelevant, redundant or useless features. The emergence of these redundant features, not only increase the dimension of the feature space, reducing the efficiency of learning, but also increases the possibility of noise data, and thus interfere with learning, the learning process mining algorithms, and ultimately affect the classification model structure. This has been confirmed by many studies, such as the Langley [2] pointed out that the number of samples nearest neighbor algorithm and its computational complexity is the number of irrelevant features with exponentially growing. In addition, the decision tree algorithm in the logical XOR condition, the required sample complexity is not related to the number of features with exponentially growing; Bayesian classifier prediction performance characteristics are more sensitive for redundancy. Therefore, to reduce such adverse

factors and reduce the dimension of the feature space, the characteristics of irrelevant or redundant data should be removed to reduce the interference noise data, and effectively improve the efficiency and performance of the learning algorithm, and to avoid the number of samples occurs when there are fewer over-fitting (Over-fitting) phenomenon.

Dimension reduction, generally in two ways for high-dimensional data dimension reduction, which is the Feature extraction, Feature extraction, and Feature selection, Feature selection) [3]. Feature extraction is also known as Feature conversion, (Feature transformation) or characteristic structure (Feature construction), it by mapping or transformation methods such as the data of high dimensional Feature space is transformed into low dimensional space said process. Secondary characteristics, namely the characteristics of low dimensional space is obtained after mapping characteristics, they are usually linear or nonlinear characteristic of the original composition. This can be seen, feature extraction is based on the original feature, constructed by the combination of a new low-dimensional feature space, so that these features can better expression characteristics of the data, and the learning algorithm is trained on the characteristics of these combinations in order to obtain better learning training effect. Typical feature extraction methods include principal component analysis, independent component analysis, factor analysis, linear decision analysis, and singular value decomposition, etc.

Filter selection model is characterized by its specific classification algorithms and are independent and can be used alone as a preprocessing step for classification learning algorithm. This is particularly advantageous under certain circumstances, such as large-scale data processing or online data. Generally, Filter model mainly through the intrinsic characteristics of the sample data itself evaluate the degree of importance characteristics, such as the statistical correlation coefficient, mutual information and Fisher scores and so on. Liu et al [4] the existing Filter model evaluation criteria are divided into distance standards, conformance criteria , standards and information standards dependence of these four categories. For example, Relief [5] and its variants Relief F and I Relief such as the Euclidean distance are used to measure the degree of importance of feature subset. Dash and Liu [6] using the consistency factor in distinguishing samples to measure characteristics of the discriminant capability. Wei and Billings [7] evaluated using the squared correlation coefficient of each feature in distinguishing different classes played by the degree of importance. Abe and Kudo [8] using bayes error boundary to select the category related features.

With the other three metrics is different, information standards is through the concept of entropy in information theory to quantify the degree of uncertainty among features. Because they do not require prior data distribution is assumed known and can effectively measure the nonlinear relationship between features, information measure attracts attention. Many have been proposed based on information entropy feature selection algorithm, such as Yu and Liu [9] using symmetric uncertainty measure the correlation between features and redundancy is a typical representative. Peng and Ding [10] in mRMR selection algorithm is put forward using mutual information to estimate the candidate features and classification categories and the correlation between the selected features and redundancy, including mRMR choose with the most relevant and every time with the selected features of minimum redundancy. Bell and Wang [11] the uncertainty coefficient as evaluation standards they choose features, only when the candidate feature can bring more information to the selected feature growth degree, they may be chosen.

Wrapper feature selection algorithm as the model will be an integral part of the learning algorithm, and the direct use of the classification performance as a characteristic degree of importance of the evaluation criteria. It is based on the selected subset will eventually be used to construct a classification model , so if the classification model is constructed , the direct use of those who can achieve a higher classification performance characteristics can thus obtain a higher classification performance classification model . For example, Guyon, etc [12].The support vector machine (SVM) classification performance as feature selection evaluation criteria, proposed a backward elimination feature selection algorithm SVM-RFE. In addition to direct use of the classification performance as the evaluation criteria, some literature focus feature subset generation problem , such as the use of genetic algorithm (GA) or evolutionary algorithm heuristic method to get a subset of better performance , in order to avoid falling into local optimization problem . Typical examples such as Huang et al [13] using a hybrid genetic algorithm together with the classification for feature subset, and ultimately significantly improve the classification performance. Filter and Wrapper models all have their own advantages and disadvantages , such as high efficiency Filter selection algorithm to obtain a subset of the suit different learning algorithms , but ultimately performance is not high ; Wrapper although able to get higher performance, but the algorithm itself is less efficient, and prone to over-fitting phenomenon .

Machine learning research boom also to feature selection has injected fresh blood. Such methods are characterized by simultaneous use of multiple feature selection algorithms to obtain a subset of single or multiple features to improve the performance of the final classification model. Feature selection based on machine learning algorithms can be broadly divided into two categories: serial and parallel selection method. As the name implies, a serial feature selection method using a serial way to organize multiple feature selector, namely a feature selector output is another feature selection input. Das proposed BDSFS [14] is a typical representative of this approach. It Boosting learning techniques combined with feature selection procedures , each selection feature, estimates are made of Decision Dump selected classification performance characteristics , and so updates the sample weights , in order to select the next feature . Finally the selected feature integrated to obtain a final classification model. Parallel feature selection method is the use of a variety of integrated learning organization in parallel selection algorithm, making them independent of each other without disturbing each other, so they are also known as integrated feature selection. The integrated feature selection method first feature selection algorithm to generate a plurality of feature subset or classification model, and then according to the combination strategy or model these subsets are combined to get the final result. This integrated approach parallelism mainly reflected in two aspects: one is the use of sampling techniques to generate different subsets of training samples, the other is to use different feature selection algorithm. For example , Li et al. [15] Bootstrapping first sample of the data samples , and then using the trained SVM, and calculate the corresponding AUC (Area Under Curve) value, and so determine the degree of importance of each feature , remove unimportant features , then again to obtain a plurality of SVM classifier training, and finally integrated operations , to obtain a final result[21-22].

Although the feature selection algorithm USES a variety of information measures, existing selection algorithm, the characteristics of the information entropy is stay the same throughout the selection process, this does not accurately reflect the feature selection is a process of dynamic change. To address this issue, this paper presents the concept of dynamic mutual information, and gives conditions based on dynamic mutual information and dynamic mutual information feature selection algorithm. Such selection algorithm in the calculation of the process of mutual information reference classification tree structure principle, that once the sample data can be

identified by the selected features, then it is for the purposes of unselected feature redundant or unimportant.

II. RELATED WORD

A. Feature Selection

Feature selection is statistical, machine learning and data mining problems in the field of classical studies, it is to solve the problem of large-scale data derived. Given the sample data set T= (O, F, C), wherein $F = \{ f_1, f_2, \ldots, f_m \}$ $C = \{c_1, c_2, \ldots, c_k\}$, $O = \{o_1, o_2, \ldots, o_n\}$ respectively, and characteristics, category, and data sample set. So as $J : 2^F \rightarrow [0,1]$ feature subset evaluation function, where J (X) value indicates the amount of information contained in feature subset X more. In this case, the feature selection algorithm generally have the following three types: 1) from the feature set F to find a feature subset X, such that J (X) max ; 2) a given threshold value J0, from the minimal F to find a set X, such that J (X)> J0; 3) from the F to find a subset of X, such that J (X) as large as possible , and as little as possible in X number of features . These three representation reflects the different aspects of feature selection and focus, the first of which focuses on the amount of information contained in the selected subset of features, namely the selection process as much as possible without loss of information; second emphasizes Select a satisfy a given condition minimal subset; final one is in the subset size and the amount of information to take a compromise between the values. This can be seen, the evaluation function J (X) is an important factor in feature selection, which can be expressed in various forms, such as the classification accuracy, the conditional probability distribution, or information entropy. What form used in the actual situation, but depends on the specific situation. In general, a common convention is acceptable based feature selection is given evaluation criteria, from the original feature space; select a subset of features, the most relevant to the process of the target concept.

Normally, the process of feature selection is done before training classification learn algorithm, it can be used as a preprocessing step of learning algorithm. Sample data set by the relevant pretreatment (such as collecting, lost data filled, the standard normalization, etc.) after being fed to the input parameters as feature selection algorithms. Then, the feature selection algorithm based on the evaluation criteria given, remove those features irrelevant or redundant, and retain the characteristics that satisfy the determination condition. Finally, the remaining samples and class characteristics and form a new data set, and provided to the classification learning algorithms in order to obtain the final classification predictive models.

This can be seen, feature selection algorithms for classification learning also plays an important role, because it chooses a subset of the merits of the classification model directly determines the final performance. On the one hand, a good subset of features can significantly improve the efficiency of classification learning algorithm and classification model performance, and to some extent, improve the generalization ability of classification model, thus effectively avoiding the interference of noise data. Classifier performance good or bad, on the other hand, also reflect the advantages and disadvantages of the selected subset, namely good subset, should contain classification information as much as possible. Therefore, in the classification, the feature selection algorithm that can usually selected classification performance high, and to minimize the number of feature subset. This is also the classification performance of some algorithms used directly as one of the main evaluation function.

In general, the feature subset selection process from the initial setting, the search strategy, evaluation and termination conditions subset of these four steps. The initial set is a subset of feature selection algorithm beginning as well as the starting point for the search process, the results of its choice behind search strategy has a direct impact. Typically , if the initial subset S is empty , the algorithm initially not selected characteristics, then the subsequent search process will select one by adding the candidate feature subset , which is called the forward search ; if the initial subset of the original feature space. , i.e. S = F, then the search process will select the subset S continuously remove unimportant or irrelevant features , this is called the backward search ; if the initial subset of the feature set F is randomly generated , the search process tend to adopt random search strategy to add or remove the selected candidate feature characteristics. Termination conditions are based on the candidate subset evaluation score J (S) or other constraints determine whether the current candidate subset S meet preset conditions. If the conditions are met, then the end of the selection algorithm and returns the candidate feature subset S as a final result; otherwise the search process continues to cycle, generate new candidate subset until the termination condition is satisfied. Feature selection algorithm is often used in the following termination conditions : 1) the candidate number of features in the subset S exceeds a threshold value given in advance ; 2) Search of cycles exceeds a preset threshold value ; 3) evaluation function value J (S) the highest or optimum ; 4) evaluation function value J (S) exceeds a threshold value given in advance .

Search strategies and evaluation criteria feature selection algorithm is two key issues. Choose a good search strategy can accelerate the speed to find the optimal solution; good evaluation criteria can be assured of the selected subset has a wealth of information, reduce false choices and improve the algorithm performance. Feature selection process is a subset of the search to some extent optimization problem, where each candidate subset S is a state of the search space. Optimization in the subset of the specific process, the generation of the candidate subset of the search direction is the feature selection algorithm one of the issues to be considered , it is mainly the following four forms: Forward Search: current candidate subset increase one or several new features ; backward search : a subset of the current candidates , delete one or several characteristics ; bidirectional search : a subset of the current candidates , first remove certain features, and then add a number of new features ; random search : random to generate a set of candidate sub . The first three search direction Select greedy strategy often used to add or remove a candidate feature, namely to increase the unselected feature all the best performance characteristics to the current candidate subset S, or from a candidate subset S removes the worst performance characteristics, purpose of doing so is to make the search process every step towards the best direction to expect to find the optimal solution. Evaluation criteria are mainly based on some measure of criteria for the selected feature subset of the merits of its assessment of the extent of the means. As the evaluation criteria selection algorithm directly determine the output results and classification model performance, so it is in the feature selection algorithm occupies an important position. In addition, the same selection algorithm using different metrics may produce different "optimal" feature subset. Because of this, the choice of evaluation criteria feature selection algorithm has been a research focus. So far, a number of evaluation criteria have been proposed, which can be divided into the following five: The distance metric, consistency metric, metric dependent, information classification error metrics and metric. Information metrics quantify the main features of the use of information entropy relative to the degree of uncertainty classification categories to determine the content it contains classified information. Measure information advantage is that it is a kind of no parameters, nonlinear standard, and it does not need to know the distribution of the sample data in advance. Since the information entropy can be well quantified characteristics relative to the category of the degree of uncertainty, so it is in the feature selection algorithm has been widely concerned.

B. Mutual information and conditional mutual information

Information is an objective state of affairs and sports a universal form, which is to ensure that the objective world or the system has a certain internal structure and functionality. The objective world, there are all kinds of news and information is the message contains the new knowledge or new content, used to enhance their awareness of objective things, thereby reducing the uncertainty of knowledge.

Mutual information (Mutual information) to measure the strength of two variables, the degree of interdependence between the introduction, which represents two variables jointly owned information content. Given two random variables X and Y, if they are on the verge of a probability distribution respectively, p (X) and p (Y), and then the mutual information between them I (X, Y) is defined as:

$$I(X;Y) = -\int_y \int_x p(x,y) \log \frac{p(x,y)}{p(x)p(y)} dxdy \qquad (1)$$

P (x, y) is the random variables X and Y of the joint probability distribution. By definition, when the variables X and Y is completely unrelated or independent of each other, their mutual information is zero, to a minimum, which means they do not exist between the same information; Conversely, when they are higher the degree of interdependence , the mutual information I (X; Y) value greater, the same information contained in the more.

Conditional mutual information is given under the condition of a random variable, the other degree of interdependence between the two variables. In other words, it is the expression of a situation occurs in the known case of the other things, the degree of association between. If the random variable Z is known, so variables X and Y about Z conditional mutual information for:

$$I(X;Y \mid Z) = -\int_z \int_y \int_x p(x,y,z) \log \frac{p(x,y \mid z)}{p(x \mid z)p(y \mid z)} dxdydz \qquad (2)$$

The p (x, y, z) is a joint probability distribution of p (x | z), p (y | z) and p (x, y | z) are the conditional probability distribution. By mutual information, and entropy definition, formula (2) can be represented as the following equivalent form:

$$\begin{aligned} I(X;Y \mid Z) &= H(X \mid Z) - H(X \mid Y,Z) = H(Y \mid Z) - H(Y \mid X,Z) \\ &= H(X \mid Z) + H(Y \mid Z) - H(X,Y \mid Z) \end{aligned} \qquad (3)$$

Formula (3) indicate a variable X or Y) to another variable Y (X) or how much information, and this information for other variables Z is unknown.

By mutual information, the definition, the probability distribution of random variables must be known in advance. However, real-world applications, the true probability distribution of the data is generally unknown. Therefore, the calculation of entropy or mutual information, you must first approximate the probability density distribution of the random variable. In this paper, the nature of the Gaussian kernel function approximated probability density distribution of the variables. Specifically, the Gaussian kernel function method to estimate the probability density function is:

$$\hat{p}(x) = \frac{1}{n}\sum_{i=1}^{n} G(x - x_{i,} \sum\ ) \tag{4}$$

Where G (z, Σ) is Gaussian kernel function, namely

$$G(z, \sum\ ) = \frac{1}{(2\pi h)^{m/2} |\sum|^{1/2}} \exp(-\frac{z^T \sum^{-1} z}{2h^2}) \tag{5}$$

The nature of the Gaussian kernel shows that variables X1 and X2 joint probability distribution:

$$\hat{p}(x_1, x_2) = (\frac{1}{n})\sum_{i=1}^{n}\sum_{j=1}^{n} G(x_i - x_{j,} \sum_{1} + \sum_{2}\ ) \tag{6}$$

Wherein the $\sum i$ variable covariance matrix. They often take the same values. To avoid the probability density estimation requires a lot of time, Huang and Chow [16] calculated directly using the Gaussian kernel mutual information to quickly approximate quadratic mutual information,

$$I(X_1, X_2) = \log \frac{V_{(X_1, X_2)^2} . V_{(X_1)^2} . V_{(X_2)^2}}{V^2_{(X_1 X_2)}} \tag{7}$$

Of which:

$V_{(X_1, X_2)^2} = \sum_{i=1}^{n}\sum_{j=1}^{n}\prod_{k=1,2} p(x_i)p(x_j)G(x_{1i} - x_{2j}, 2\sum\ )$;

$V_{(X_k)^2} = \sum_{i=1}^{n} p(x_i)\sum_{j=1}^{n} p(x_j)G(x_{1i} - x_{2j}, 2\sum\ )$;

$V_{(X_1 X_2)} = \sum_{i=1}^{n} p(x_i)\prod_{k=1,2}(\sum_{i=1}^{n} p(x_j)G(x_{1i} - x_{2j}, 2\sum\ ))$;

III. FEATURE SELECTION BASED ON CONDITIONAL DYNAMIC MUTUAL INFORMATION

Seen from the foregoing discussion, the main objective of feature selection from the sample data set T = (O, F, C) of the original features F to find a subset of S, such that it contains as much information about the class distinction, which contain more and Category C -related knowledge, but also makes the degree of redundancy within the subset as small as possible. Feature subset S ability to distinguish between the classes can be represented in a way and Class C correlation between the degree of dependence and feature subset S of class distinction between the stronger , S and C dependent on the degree of correlation between the higher . Feature subset S degree of redundancy as far as possible little mean contains different information between the selected features, namely, each of the selected features are important, indispensable. It also shows that from the side feature subset S is the number of features included should be minimal, because the more the number of selected features, S may be higher degree of redundancy. For most of the information measure as an evaluation criterion feature selection algorithm, even though they use different metrics information, and the manifestations vary widely, but they all follow a common choice framework, and information can also be generalized metrics expressed in the form.

Through in-depth study found that the current proposed selection algorithm based on information measures are calculated in the whole sample space characterized by information entropy, which features information entropy in the entire selection process no change , because once a given sample data set after they are fixed. This is obviously not reasonable, because it does not reflect the feature selection is a dynamic process, we can't accurately measure the specific characteristics of the selection process interdependencies between the various levels. As feature selection process continues, the data set the sample data will continue to be identified has been selected characteristics, i.e. they can be a subset of the selected feature for classification, while the number of samples can't be identified less and less. Since these can be recognized in the sample data relative to those who have not yet been selected features a redundant or useless, so they can be in the feature selection process removed from the sample data set. In this case, the original estimate in the whole sample space information entropy can't truly reflect this characteristic. To address this problem, this chapter proposed the concept of dynamic mutual information, which is not recognized on a sample re- valuation. On this basis, the dynamic conditions of mutual information (CDMI) feature selection algorithm.

A. dynamic mutual information and conditional dynamic mutual information

Feature selection algorithm is the information in either metric or other metrics, which are usually measured feature or subset of the degree of correlation between categories. Correlation with the probability defined similar information entropy or mutual information is also often used to describe or measure the degree of correlation between features. Information metrics to quantify the advantage is that it can accurately describe the characteristics in the form of the degree of uncertainty. Because of this, the information has been widely used measure of feature selection algorithm, and its performance has been confirmed by many experiments, as such. Although the information presented metrics have their different representations, but according to the entropy or mutual information definition, they are built on the basis of probability theory. Thus, the degree of correlation metric characteristics, whether it is information or probability metrics metric , which are characterized in the sample needs to be calculated in advance the probability distribution dataset situation . Noted that once a given sample data set , the features in this sample space probability distribution is determined, down, and across the feature selection process will no longer be changed. This does not change the situation will generate a new problem, that metric does not accurately reflect the information or uncertainty of dynamic change, because feature selection is a dynamic process. This dynamic process is characterized in that has been selected with the increase in the uncertainty of class C is gradually reduced, while the sample space can't identify the number of samples also showed a decreasing trend. This description does not change in a way the correlation metrics contain some "false" information.

From the perspective of the learning algorithm, the sample data classification can generally be divided into two types: the identified samples and samples not recognized. As learning process continues, sample concentration did not identify the data being used to study or training, whether to tags can be accurately identified, the quantity is less and less. When the sample concentration unidentified samples, the learning process should continue to be unable to identify a sample of data classification learning how to operate until all of the samples can be correctly identified so far. Feature selection process is similar, with the continuous selection candidate feature, C degree of uncertainty decreases. The entire feature selection process will be terminated when the selected feature subset of the information content and the amount of information approximate original feature space. This means that all the samples when the sample concentration can be selected subset of features identified, the feature selection process will end; Conversely, if there are still unidentified sample set of samples, then the selection process will continue to execute.

B. the proposed algorithm

Assumes the currently selected feature subset is S, sample set T is correspondingly divided into two disjoint parts: the identified samples O_l and unidentified samples O_u, feature selection process every step of the candidate feature set F from selecting a candidate S is added to the feature f. If the candidate feature f can't recognize a sample O_u as much as possible into the sample identification, then it is a good feature, at which point it will be preferred feature selection process. Noting that the existing selection algorithm selection process every step of the generally choose the candidate with the largest mutual information feature f as a reference object. However, if we let S and F represent the currently selected feature subset and candidate subset, then for a sample O_l is identified, any candidate feature $f \in F$ is for Class C are irrelevant or redundant.

Based on the above discussion, the following gives a new feature selection algorithm, which uses the dynamic conditions of mutual information as a feature selection metrics. Dynamic mutual information because the condition is not recognized on the sample estimate, so the selection process every step of the need to preserve those characteristics can be selected by the correct identification of the sample information, and remove them from the sample set to ensure that the candidate feature mutual information value can be accurately estimated. Specific algorithm implementation details are as follows:

Input: A training dataset T = (O, F, C);

Output: A feature subset S;

Initialize relative parameters, e.g., S = Φ, Ol=Φ, where Ol denotes the set of samples recognized by features;

While |F|≠0 and |O|≠0 do

For each candidate feature f in F do

Calculate the mutual information I(C; f | S) of f with C;

If I(C; f | S) =0, then remove it from F;

Select the feature f with the maximal I(C; f | S), and insert it into S and remove it from F, i.e., S = S + {f}, F = F – {f};

Obtain unrecognized samples by f and save them to Ol

Remove the samples in Ol from O;

End while;

Return the feature subset S as the selected subset;

The algorithm works is relatively simple, it is first calculated for each candidate feature f in F and Class C mutual information I (C; f | S). If the mutual information I (C; f | S) is 0, then the feature f will be removed directly, and class C because it is totally irrelevant. That is to say, did not identify the sample data for f is completely random distribution, in this case the candidate features for classification categories of prediction is f house with no contribution. Mutual Information valuation will be sorted in descending order. Candidate for the highest value of mutual information feature f will prefer and add to the selected subset of S. Subsequently, the algorithm to obtain a sample can be identified by feature f O_l. Since these candidates for the other features of the sample data is redundant, so that they will delete the original sample set O , O retain only those that have not been correctly recognized sample data. The aim is to ensure that in the subsequent selection process, characterized by mutual information and the ability to accurately measure the candidate feature the degree of correlation between categories. Sample after deleting O_l are identified, the selection process will choose other candidates into the next round of cycle characteristics. This process loops until the candidate feature set is empty, or all the sample data can be correctly identified so far.

IV. EXPERIMENTAL COMPARISONS

To validate the performance of feature selection algorithm the algorithm in this section feature selection algorithm with other data sets in the UCI test simulation experiments to compare their performance. Experiments in the use of other information-based measure five typical feature selection algorithm as an experimental comparison object: BIF [17], MIFS-U [18], FCBF [9], MIFS [19] and mMIFS-U [20], where the uncertainty FCBF SU using symmetric characteristic correlation measure. As mentioned earlier, these five selection algorithms were used five different metrics evaluated characteristics of the information.

A. Dataset

In order to fully compare these six feature selection algorithm performance, simulation experiments using 11 different sizes UCI common test data sets. All these sample data sets from the UCI Machine Learning repository, they are often used to compare the field of machine learning and data mining algorithms or learning feature selection algorithm. Figure 1 gives a brief summary of these test data set information, such as name, number of samples, the number of features and the number of categories, etc. For further description of the dataset can refer to UCI machine learning sites. From the data in Figure 1 it can be seen, these data sets contain a different

number of sample data, characteristics and classification categories, wherein the number of samples in the range of between 355 and 8124, the number of features is from 22-1558. The number of categories corresponding to the number of classification categories, including the value "2" indicates that the data sets in data classification problem is a binary classification problem, while others value it means that the data set is a multi-classification problems.

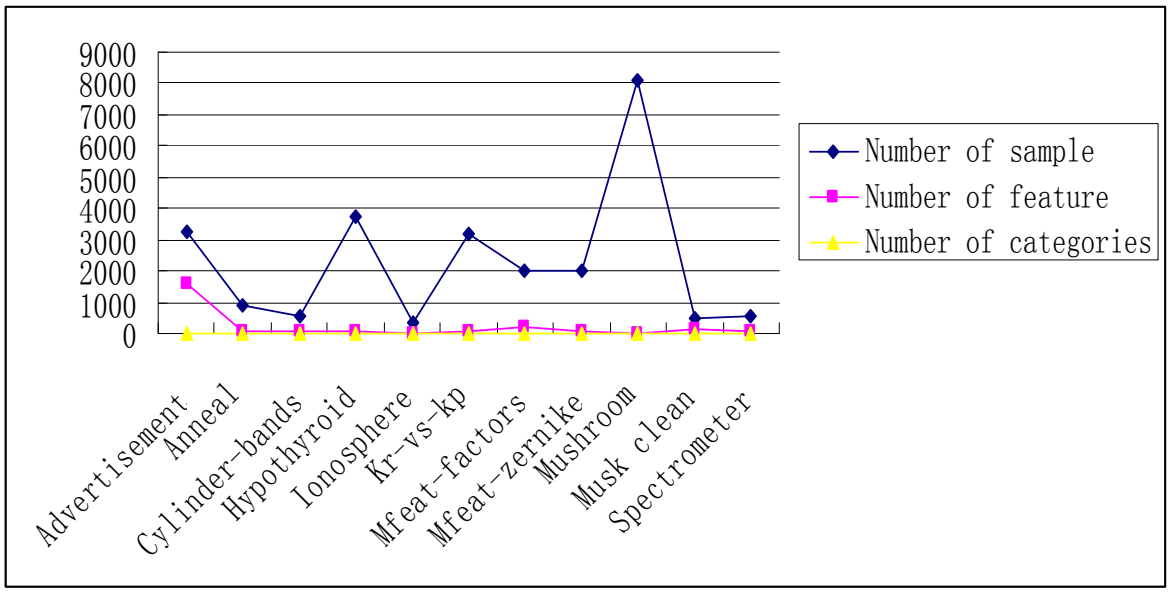

Figure 1. Summary of experimental data sets described

Because all of these data sets from real specific application areas, so some features are too specific reason that is not suitable for classification. Such features will be removed before simulation, to prevent the classification model of structure fitting occurred.

B. Setting

Sample data set after the end of the pretreatment, feature selection algorithm as input parameters for feature selection operation. In order to ensure the fairness of the comparison experiments, each feature selection algorithm will select the same number of features. Because this algorithm USES the consistency factor loop termination conditions, so will choose a feature subset, and several other feature selection algorithm for feature. Experiments using feature selection algorithms belong to the Filter selection model, with their specific learning algorithm are independent of each other. Therefore the experiment process need external learning algorithm to participate in the performance verification, to test the classification performance of the selected subset. In addition, individual learning algorithms have preference is likely to be some kind of feature selection algorithm, namely the learning algorithm may produce better performance, the

results of some kind of selection algorithm classification model caused by the over fitting phenomenon. Therefore, simulation experiments also USES two typical external classification learning algorithm: nearest neighbor (nn) algorithm and decision tree C4.5 algorithm. Among them 1 nn is learning method based on the samples; The C4.5 decision tree method is the typical representative. Choose this two kinds of learning algorithm is the main reason is that their learning efficiency is relatively high, and they are integrated in the data mining Weka1 software. In the process of concrete experiment, the learning algorithm of the parameters will be set to the default value.

In order to obtain more reliable results, take 10 times cross-validation experiments verify the classification performance mode, and repeat three times, and finally take the average as the final result. That is, the learning algorithm runs on each test data set three times, each time using 10-fold cross validation method, the final result is the average of these three.

C. Experiments and analysis

1. Select the number of features

This selection algorithm is given in Table 1 CDMI data sets in each of the selected number of features, including "characteristic number" column indicates the number of features selected algorithm, and the "rate" column indicates the number of features selected ratio of total number of features with the original. As can be seen from the table, CDMI algorithm in most cases (except the outer Kr-vs-kp) can remove most of irrelevant features, leaving only a small part of important features. For example, Mfeat-factors dataset only original features 216 selects one of eight important features; while Advertisement CDMI data sets selected number of features of the original amount of 7%.

Table 1 CDMI select the number of features

	Number of Selected feature	Radio(%)
1	108	6.92
2	8	21.05
3	14	36.23
4	14	48.28
5	9	23.89
6	31	84.01
7	9	3.82
8	13	26.18
9	5	19.26
10	18	11.36
11	27	27.01

2. Separate classification performance

Figure 2 and Figure 3, respectively, and C4.5 1NN two learning algorithms feature selection algorithm using the results before and after the classification, where the "raw" column indicates the classification learning algorithm feature selection algorithm is not used in the case of classification performance, the study algorithm on the original data set classification accuracy; tables in " mean " represents a different feature selection algorithm in all data sets the average performance. Bold values in the table represent the value of these six feature selection algorithm is the highest. From Figure 2 it can be seen in the classification performance, CDMI selection algorithm in six data sets on the classification performance is the highest, this figure more than several other selection algorithms, such as mMIFS-U algorithm performance is only 2 max; except in Mfeat-factors significantly reduced data set classification algorithm performance outside, CDMI in the remaining 10 data sets even degrades performance, but reduces the magnitude is not high. In fact, several other selection algorithm Mfeat-factors are also significantly reduced the performance of classification models, which may be due to the small number of the selected characteristic caused by this reason. Further, CDMI average performance of algorithm, though lower than in the original space 1NN value obtained, but it is the highest in the other selection algorithm.

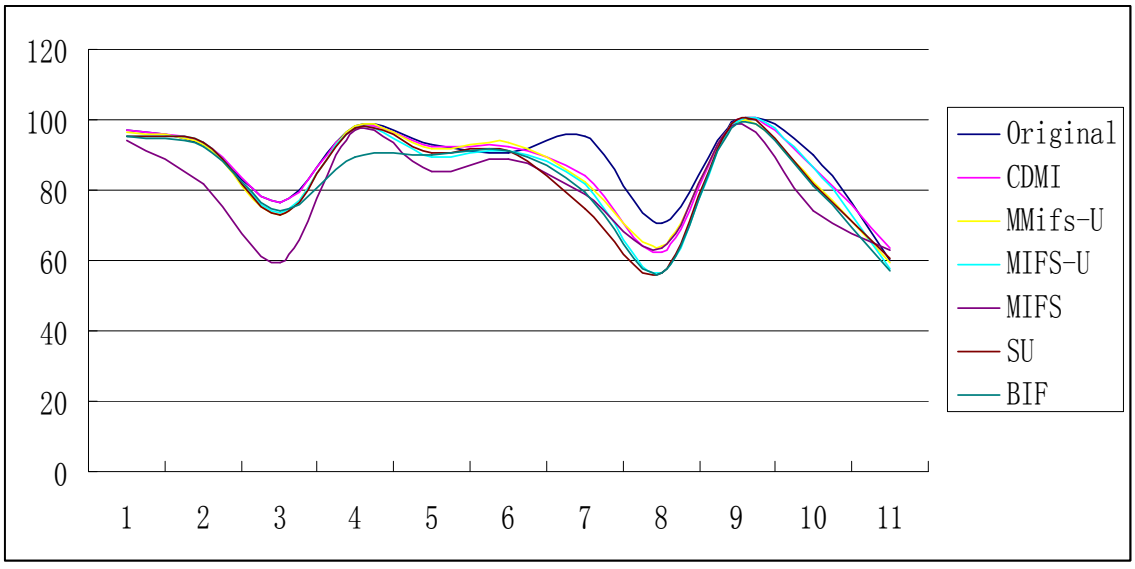

Figure 2. six kinds of feature selection algorithm in the classification 1NN classifier performance comparison (%)

In C4.5 learning algorithm (shown in Figure 3) and, CDMI algorithm performance advantage, although not so obvious in the 1NN, but in comparison, several options are still better than the other algorithms, such as, CDMI maximum performance of the algorithm the number is 4, and MIFS, SU and BIF number three algorithms are 2,2 and 1, respectively, and it is in these 11 data sets performance compared with other algorithms no one is the worst, and the maximum the number and value of the average performance, CDMI are optimal.

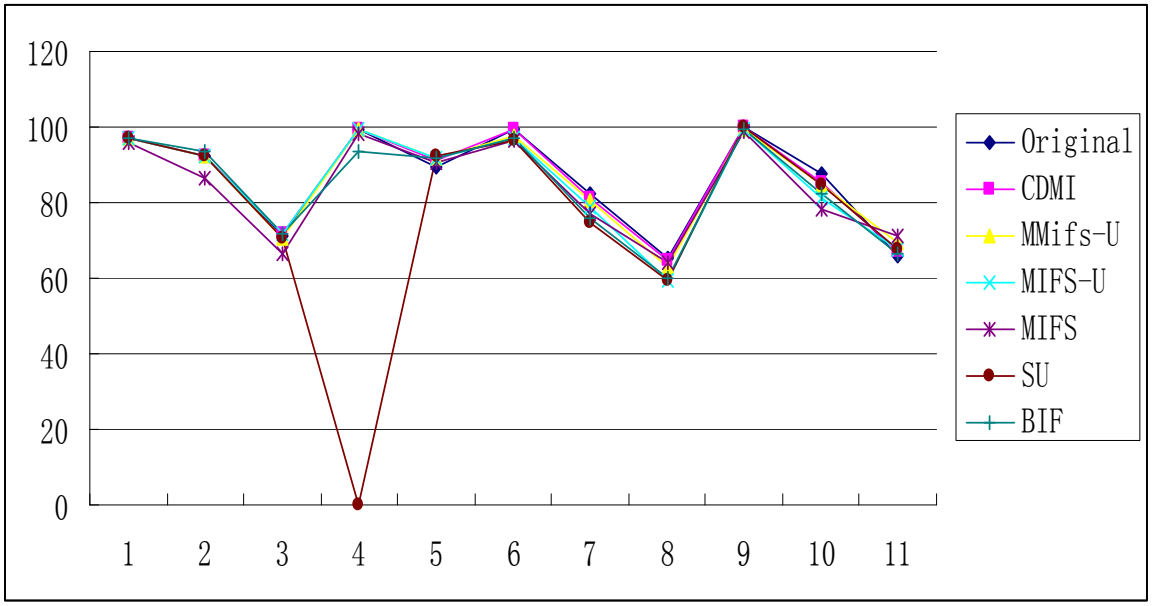

Figure 3 six kinds of feature selection algorithm C4.5 classifiers in the classification performance comparison

3. Determination of key point directions

Because each learning algorithm has its own characteristics or advantages, so the performance of a single classification model is good or bad, and not enough selection algorithm performance advantages and disadvantages. To describe the different options from the overall performance of the algorithm, we will 1NN and C4.5 these two learning algorithms on each data set averaged sum of classification accuracy, the final result as shown in Figure 4. As can be seen from the table, CDMI performance on the algorithm to the same selection algorithm is better than the other types. For example, CDMI algorithms in 11 datasets maximum performance on the number 5; CDMI is the average performance of these six selection algorithm in the highest.

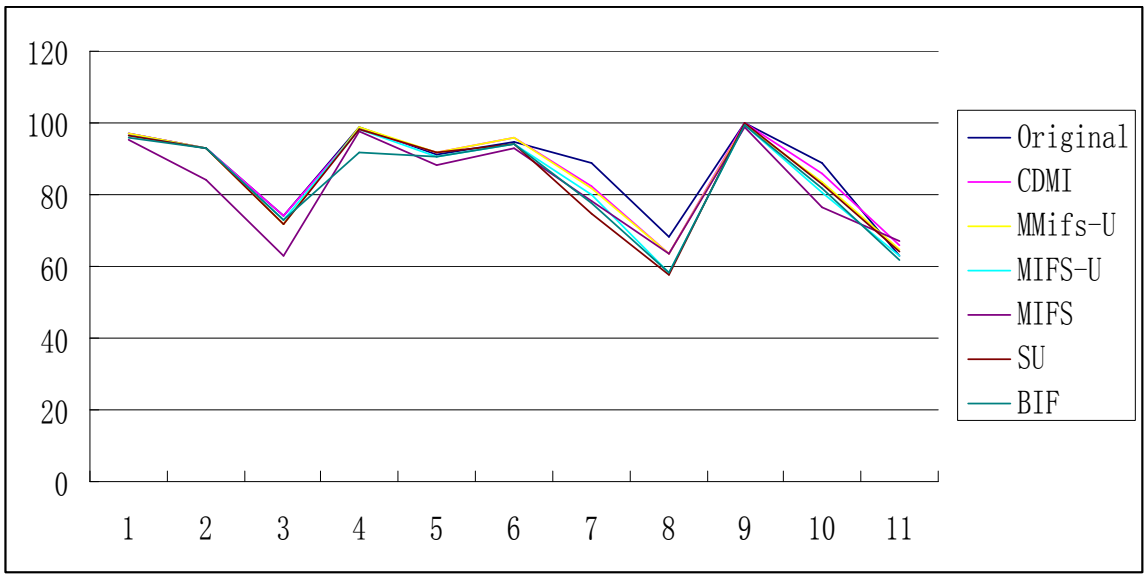

Figure 4. six kinds of feature selection algorithm, the average classification performance comparison

In addition, we also their average performance for statistical t-test analysis to further verify the selection algorithm exists between significant differences. Bar 1 show the five other CDMI selection algorithms with the average performance of the statistical t-test comparison result, where the horizontal axis represents the data set, the vertical axis represents the p-value of t-test. If the bar is located in a horizontal line in the figure 5 (value of 0) above, it means that the selection algorithm is better than the average performance benchmark selection algorithm. Conversely, located below the horizontal bar indicates that the selection algorithm is worse than the baseline algorithm. If the p-value is greater than the absolute value of 2 indicates that the selection algorithm performs significantly better or worse than the baseline algorithm (95% confidence).

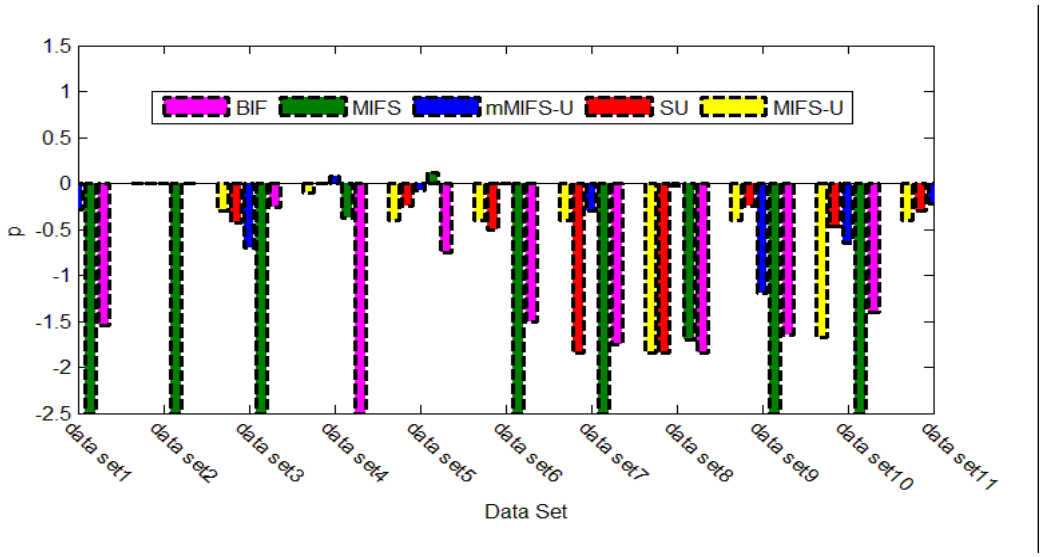

Figure 5 CDMI selection algorithm classification performances with other statistical t-test comparison

V. CONCLUSIONS

Currently most of the selection algorithm for the selection process in the specific entropy or mutual information between features valuation does not accurately reflect the degree of problems related to, this chapter also proposed the concept of mutual information dynamic conditions, the valuation of mutual information that is not in the selection process identification of the sample space, rather than for the entire sample space. On this basis, given the dynamic conditions of mutual information based feature selection algorithm. In order to verify the proposed selection algorithm performance, CDMI and the five other typical metrics based on information feature selection algorithm in the 11 UCI data sets on a common test simulations to compare. Experimental results show that, CDMI selection algorithm performance in most cases the performance is better than the other five selection algorithm. It can be seen from the experimental results, CDMI data algorithm sensitive to noise, such as Kr-vs-kp selected on some redundant features. So the next step is to use the other end of the main work conditions or ways to avoid the interference of noise data.

V. ACKNOWLEDGEMENT

The work was supported by the fund of Collaborative Innovation Center of Tongling University.

REFERENCES

[1] Fayyad U, Piatetsky-Shapiro G, Smyth P. From Data Mining to Knowledge Discovery in Databases, AI Magazine, vol.17, pp. 37-24, 1996.

[2] Langley P. Selection of relevant features in machine learning, Proc of the AAAI Fall Symposium on Relevance, Menlo Park, pp.140-144, 1994.

[3] Song Q, Ni J, Wang G. A fast clustering-based feature subset selection algorithm for high-dimensional data, Knowledge and Data Engineering, IEEE Transactions on, vol.25, no.1, pp.1-14, 2013.

[4] Lopez M I, Luna J M, Romero C, et al. Classification via Clustering for Predicting Final Marks Based on Student Participation in Forums, 5th International Conference on Educational Data Mining, pp. 148-151, 2012.

[5] Kira K, Rendell L. A practical approach to feature selection, Proc of the 9th International Conference on Machine Learning, pp. 249-256, 1992.

[6] Dash M, Liu H. Consistency-based search in feature selection, Artificial Intelligence, vol. 151, no.1-2, pp. 155-176, 2003.

[7] Wei H-L, Billings S A. Feature Subset Selection and Ranking for Data Dimensionality Reduction, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.29, no1. 162-166, 2007.

[8] Yin L, Ge Y, Xiao K, et al. Feature selection for high-dimensional imbalanced data, Neurocomputing, vol. 105, pp. 3-11, 2013.

[9] Li B Q, Hu L L, Chen L, et al, Prediction of protein domain with mRMR feature selection and analysis, PLoS One, 2012, 7(6): http://www.plosone.org/article/info:doi/10.1371/journal.pone.0039308.

[10] Peng H, Long F, Ding C. Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.27, no.8, pp. 1226-1238, 2005.

[11] Foithong S, Pinngern O, Attachoo B. Feature subset selection wrapper based on mutual information and rough sets, Expert Systems with Applications, vol.39, no.1, pp. 574-584, 2012.

[12] Guyon I, Weston J, Barnhill S, et al. Gene selection for cancer classification using support vector machines, Machine Learning, vol.46, no.1-3, pp.389-422, 2002.

[13] Suzuki T, Sugiyama M. Sufficient dimension reduction via squared-loss mutual information estimation, Neural computation, vol.25, no.3, pp. 725-758, 2013.

[14] Das S. Filters, Wrappers and a Boosting-Based Hybrid for Feature Selection, Proc of the 18th International Conference on Machine Learning, San Francisco, CA, USA, pp. 74-81, 2001.

[15] Li G-Z, Meng H-H, Lu W-C, et al. Asymmetric bagging and feature selection for activities prediction of drug molecules, BMC Bioinformatics, vol.9, no.S6, pp. 7-11, 2008.

[16] Huang D, Chow T W S. Effective feature selection scheme using mutual information, Neurocomputing, vol.63, pp. 325-343, 2005.

[17]Jain A K, Duin R P W, Mao J. Statistical Pattern Recognition: A Review, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.22, no.1, pp. 4 – 37, 2000.

[18]N. K. Suryadevara and S. C. Mukhopadhyay, "Determining Wellness Through An Ambient Assisted Living Environment", IEEE Intelligent Systems, May/June 2014, pp. 30-37.

[19]Foithong S, Pinngern O, Attachoo B. Feature subset selection wrapper based on mutual information and rough sets, Expert Systems with Applications, vol.39, no.1, pp. 574-584, 2012.

[20]N. K. Suryadevara, S. C. Mukhopadhyay, R.Wang, R.K. Rayudu and Y. M. Huang, Reliable Measurement of Wireless Sensor Network Data for Forecasting Wellness of Elderly at Smart Home, Proceedings of IEEE I2MTC 2013 conference, IEEE Catalog number CFP13IMT-CDR, ISBN 978-1-4673-4622-1, May 6-9, 2013, Minneapolis, USA, pp. 16-21.

[21]Ang K K, Chin Z Y, Zhang H, et al. Mutual information-based selection of optimal spatial–temporal patterns for single-trial EEG-based BCIs, Pattern Recognition, vol.45, no.6, pp. 2137-2144, 2012.

[22]A. Gaddam, S.C. Mukhopadhyay and G. Sen Gupta, Necessity of a Bed Sensor in a Smart Digital Home to Care for Elder-people, Proceedings of the 2008 IEEE Sensors conference, Lecce, Italy, October 26-28, 2008, page 1340-1343.

[23]Novovicova J, Somol P, Haindl M, et al. Conditional Mutual Information Based Feature Selection for Classification Task, Proc of the 12th Iberoamericann Congress on Pattern Recognition, pp. 417-426, 2007.

[24]N.K. Suryadevara, S.C. Mukhopadhyay, R. Wang, R.K. Rayudu, Forecasting the behavior of an elderly using wireless sensors data in a smart home, Engineering Applications of Artificial Intelligence, Volume 26, Issue 10, November 2013, Pages 2641-2652, ISSN 0952-1976, http://dx.doi.org/10.1016/j.engappai.2013.08.004.

[25]Daode Zhang et al., Research on Chip Defect Extraction based on Image-Matching, International Journal on Smart Sensing and Intelligent Systems, vol. 7, no. 1, pp.321 – 336, 2014.

[26]N.K.Suryadevara, A. Gaddam, R.K.Rayudu and S.C. Mukhopadhyay, "Wireless Sensors Network based safe Home to care Elderly People: Behaviour Detection", Sens. Actuators A: Phys. (2012), doi:10.1016/j.sna.2012.03.020, Volume 186, 2012, pp. 277 – 283.

[27] Yanmin LUO, Peizhong LIU and Minghong LIAO, An Artificial Immune Network Clustering Algorithm for Mangroves Remote Sensing Iamge,International Journal on Smart Sensing and Intelligent Systems, vol. 7, no. 1, pp. 116 – 134, 2014.