**Indonesian Journal of Electrical Engineering and Computer Science**
Vol. 7, No. 1, July 2017, pp. 214 ~ 225
DOI: 10.11591/ijeecs.v7.i1.pp214-225


# A Hybrid Feature Selection based on Mutual Information and Genetic Algorithm

**Yuan-Dong Lan**
School of Information Science and Technology, Huizhou University, Huizhou 516007, China
Corresponding author, e-mail: yd_lan@foxmail.com

***Abstract***

*Feature selection aims to choose an optimal subset of features that are necessary and sufficient to improve the generalization performance and the running efficiency of the learning algorithm. To get the optimal subset in the feature selection process, a hybrid feature selection based on mutual information and genetic algorithm is proposed in this paper. In order to make full use of the advantages of filter and wrapper model, the algorithm is divided into two phases: the filter phase and the wrapper phase. In the filter phase, this algorithm first uses the mutual information to sort the feature, and provides the heuristic information for the subsequent genetic algorithm, to accelerate the search process of the genetic algorithm. In the wrapper phase, using the genetic algorithm as the search strategy, considering the performance of the classifier and dimension of subset as an evaluation criterion, search the best subset of features. Experimental results on benchmark datasets show that the proposed algorithm has higher classification accuracy and smaller feature dimension, and its running time is less than the time of using genetic algorithm.*

***Keywords****: feature selection, mutual information, genetic algorithm, filter, wrapper*



## 1. Introduction

High dimensional data are often occurring when deal with images, videos and so on. It will produce large time and energy overheads when the high dimensional data is directly used in classification tasks. If no efficient feature selection is done on the data, namely without dimensionality reduction, many original classification methods would perform very badly, especially large time and energy overheads might be required [1]. Feature selection is very important in pattern recognition [2], data mining and in many other machine learning tasks [3-6], especially when dealing with high dimensional data. It removes the irrelevant and redundant features of the samples according to the given criteria, to reduce the dimension of feature space, and then chooses the most effective features for the classification [7, 8].

Feature selection method can be divided into three categories: filter method (filter), wrapper method (wrapper) and embedded method (embedded) [9]. Filter method analyse intrinsic properties of data, it does not rely on any learning algorithm, the running efficiency is high, and so it is suitable for large-scale data sets [10]. The wrapper method is to introduce a kind of learning algorithm (classifier), according to the actual effect of the classification to evaluate the effectiveness of the selected feature subset, and therefore get a higher accuracy [11]. But the Wrapper method needs to train a classifier, and the computational complexity is very high, which is not suitable for large scale data. Embedded method try to combine the advantages of both filter and wrapper methods. In Embedded method, the classifier takes advantage of its own feature selection algorithm. So, embedded method needs to know preliminary what a good selection is, which limits their exploitation. Literature [6, 7] combine filter and wrapper methods for feature selection, these combination methods can make full use of the results of filter method, to accelerate the convergence of the wrapper algorithm, and can produce an optimal feature subset that are necessary and sufficient to a high accuracy classifier.

Mutual information method is a kind of filter feature selection method based on correlation. It uses mutual information to measure the correlation between features and categories. It is characterized by high speed, but the obtained feature subset may not be optimal, and the classification accuracy is low. Genetic algorithm (GA), as an adaptive global

*Received March 28, 2017; Revised June 2, 2017; Accepted June 18, 2017*

search method, has the characteristics of parallelism and is suitable for solving multi objective optimization problems. In order to make use of the advantages of GA and improve the running speed of wrapper method based on GA on high dimensional data, we combine the filter method with wrapper method; propose a hybrid feature selection based on mutual information and GA in this paper (HFS-MIGA). (HFS-MIGA) has two phases: filter phase and wrapper phase. The filter phase removes some redundant and irrelevant features and uses the feature estimation as the heuristic information to guide GA [12]. We adopt mutual information to get feature estimation [13]. In wrapper phase is running a GA based wrapper selector. The estimated feature obtained from the filter phase is used for guiding the initialization of the population for GA.

## 2. Research Method

### 2.1. Entropy and Mutual Information

Entropy and mutual information is an important method to measure the information content of random variables [14]. Entropy is a measure of uncertainty of random variables. Assuming that X is a discrete random variable with possible values *{x_1,x_2,…,x_n }* and probability mass function P(*X*) as:

$$H(X) = E[I(X)] = E\big[-l\,n\big(P(X)\big)\big] \quad (1)$$

Here *E* is the expected value operator, and I is the information content of *X. I(X)* is itself a random variable.

The entropy can explicitly be written as:

$$\mathrm{H(X)} = \sum_{i=1}^{n} P(x_i) I(x_i) = -\sum_{i=1}^{n} P(x_i) log_b P(x_i) \quad (2)$$

Where b is the base of the logarithm used. Common values of *b* are 2, Euler's number *e*, and 10, and the unit of entropy is Shannon for *b* = 2, Nat for *b* = e, and Hartley for *b* = 10 when *b* = 2, the units of entropy are also commonly referred to as bits [15].

In the case of P($x_i$ )=0 for some i, the value of the corresponding summand 0 $\log_b$ (0) is taken to be 0, which is consistent with the limit:

$$\lim_{p \to 0_+} p log(p) = 0 \quad (3)$$

One may also define the conditional entropy of two events *X* and *Y* taking values $x_i$ and $y_j$ respectively, as:

$$\mathrm{H(X|Y)} = \sum_{i,j} p\big(x_i, y_j\big) log \frac{p(y_j)}{p(x_i, y_j)} \quad (4)$$

Where $p(x_i,y_j)$ is the probability that $X = x_i$ and $Y = y_j$. This quantity should be understood as the amount of randomness in the random variable *X* given the event *Y*.

The concept of mutual information is derived from the concept of entropy. In probability theory and information theory, the mutual information (MI) of two random variables is a measure of the mutual dependence between the two variables. More specifically, it quantifies the "amount of information" obtained about one random variable, through the other random variable. The concept of mutual information is intricately linked to that of entropy of a random variable, a fundamental notion in information theory, that defines the "amount of information" held in a random variable. Not limited to real-valued random variables like the correlation coefficient, MI is more general and determines how similar the joint distribution *p(X, Y)* is to the

products of factored marginal distribution *p(X)p(Y).* MI is the expected value of the pointwise mutual information. The most common unit of measurement of mutual information is the bit.

Formally, the mutual information of two discrete random variables X and Y can be defined as:

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) log \left( \frac{p(x,y)}{p(x)p(y)} \right) \quad (5)$$

where *p(x,y)* is the joint probability distribution function of *X* and *Y*, and *p(x)* and *p(y)* are the marginal probability distribution functions of *X* and *Y* respectively.

Intuitively, mutual information measures the information that *X* and *Y* share: it measures how much knowing one of these variables reduces uncertainty about the other. For example, if *X* and *Y* are independent, then knowing *X* does not give any information about *Y* and vice versa, so their mutual information is 0. At the other extreme, if *X* is a deterministic function of *Y* and *Y* is a deterministic function of *X* then all information conveyed by *X* is shared with *Y*: knowing *X* determines the value of Y and vice versa. As a result, in this case the mutual information is the same as the uncertainty contained in *Y* (or *X*) alone, namely the entropy of *Y* (or *X*). Moreover, this mutual information is the same as the entropy of *X* and as the entropy of *Y*.

Mutual information is a measure of the inherent dependence expressed in the joint distribution of *X* and *Y* relative to the joint distribution of *X* and *Y* under the assumption of independence. Mutual information therefore measures dependence in the following sense: *I(X;Y)*=0 if and only if *X* and *Y* are independent random variables. This is easy to see in one direction: if *X* and *Y* are independent, then *p(x,y)=p(x)p(y)*, and therefore:

$$\log \left( \frac{p(x,y)}{p(x)p(y)} \right) = log 1 = 0 \quad (6)$$

Moreover, mutual information is nonnegative (i.e. *I(X;Y)*≥0) and symmetric (i.e. *I(X;Y)=I(Y;X)).*

Generally, *I(Y;X)* should be normalized to between 0 to 1; so, we choose symmetrical uncertainty as a measure of correlation between features and the concept target, the give features corresponding weight by their symmetrical uncertainty value [16]. The feature having larger symmetrical uncertainty value gets higher weight. Symmetrical uncertainty is defined as

$$\mathrm{SU(X,Y)} = 2 \left[ \frac{I(X;Y)}{H(X)+H(Y)} \right] \quad (7)$$

### 2.2. Genetic Algorithm

Genetic algorithms (GAs) in particular became popular through the work of John Holland in the early 1970s, and particularly his book Adaptation in Natural and Artificial Systems (1975) [17]. His work originated with studies of cellular automata, conducted by Holland and his students at the University of Michigan. Holland introduced a formalized framework for predicting the quality of the next generation, known as Holland's Schema Theorem. Research in GAs remained largely theoretical until the mid-1980s, when The First International Conference on GAs was held in Pittsburgh, Pennsylvania.

GAs belong to the larger class of evolutionary algorithms (EA), which generate solutions to optimization problems using techniques inspired by natural evolution, such as inheritance, mutation, selection and crossover. A typical genetic algorithm requires:

1. a genetic representation of the solution domain,
2. a fitness function to evaluate the solution domain.

A standard representation of each candidate solution is as an array of bits. Arrays of other types and structures can be used in essentially the same way. The main property that makes these genetic representations convenient is that their parts are easily aligned due to their fixed size, which facilitates simple crossover operations. Variable length representations may also be used, but crossover implementation is more complex in this case. Tree-like representations are explored in genetic programming and graph-form representations are explored in evolutionary programming; a mix of both linear chromosomes and trees is explored

in gene expression programming. Once the genetic representation and the fitness function are defined, a GA proceeds to initialize a population of solutions and then to improve it through repetitive application of the mutation, crossover, inversion and selection operators.

There are various flavours of GAs in circulation, varying in implementation of these four parameters, but in essence the algorithms all follow a standard procedure:

1. Start with a randomly generated population of n *l*-bit strings (candidate solutions to a problem). The population size n depends on the nature of the problem, but typically contains several hundreds or thousands of possible solutions. Often, the initial population is generated randomly, allowing the entire range of possible solutions. Occasionally, the solutions may be "seeded" in areas where optimal solutions are likely to be found.
2. Calculate the fitness f(x) of each string in the population.
3. Repeat the following steps until n new strings have been created:

   Select a pair of parent strings from the current population, the probability of selection being an increasing function of fitness. Individual solutions are selected through a fitness-based process, where fitter solutions are typically more likely to be selected. Certain selection methods rate the fitness of each solution and preferentially select the best solutions. Other methods rate only a random sample of the population, as the former process may be very time-consuming.

   With the crossover probability, cross over the pair at a randomly chosen point to form two new strings. This process ultimately result in the next generation population of chromosomes that is different from the initial generation. Generally, the average fitness will have increased by this procedure for the population, since only the best organisms from the first generation are selected for breeding, along with a small proportion of less fit solutions.

   Mutate the two new strings at each locus with the mutation probability, and place the resulting strings in the new population. It is worth tuning parameters such as the mutation probability, crossover probability and population size to find reasonable settings for the problem class being worked on. A very small mutation rate may lead to genetic drift. A recombination rate that is too high may lead to premature convergence of the genetic algorithm. A mutation rate that is too high may lead to loss of good solutions, unless elitist selection is employed.
4. Replace the current population with the new population.
5. Go to step 2. This generational process is repeated until a termination condition has been reached. Common terminating conditions are:
   a. A solution is found that satisfies minimum criteria
   b. Fixed number of generations reached
   c. Allocated budget reached
   d. The highest-ranking solution's fitness is reaching or has reached a plateau such that successive iterations no longer produce better results
   e. Manual inspection
   f. Combinations of the above

Figure 1 provides a simple diagram of the iterative nature of GAs.

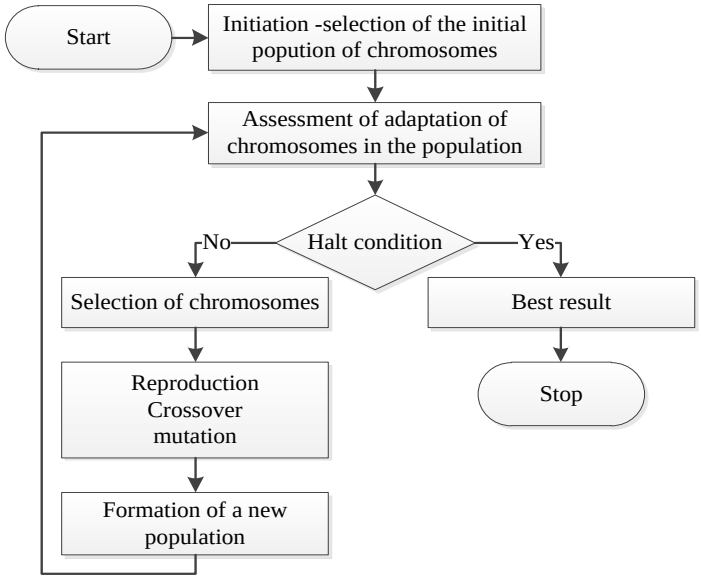


Figure 1. Flow chart of Gas

### 2.3. Framework Description of HFS-MIGA

In the first part of this paper, we introduced that the feature selection methods can divided into three categories: filter, wrapper and embedded. Filter and Wrapper each has its advantages and disadvantages, and it has strong complementarity. In this paper, we combine filter method with wrapper method, propose a hybrid feature selection algorithm based on mutual information and GA, which is called HFS-MIGA. Figure 2 provides the flow chart of HFS-MIGA.

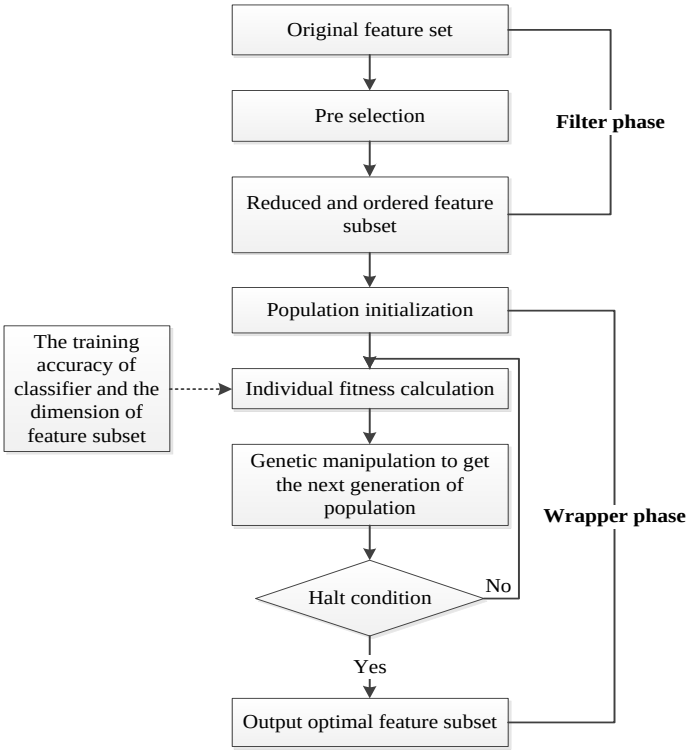


Figure 2. Flow chart of HFS-MIGA

HFS-MIGA is divided into two phases:

1) Filter phase: sort according to the amounts of mutual information between feature and category label from large to small. This filter phase has two functions.
   a. It filters out the features that are not related to the class label, and reduce the dimension of the original feature set. Especially for high dimensional data, there are often a large number of irrelevant features, filter out irrelevant features, can greatly reduce the search space of the subsequent GA, improve the efficiency of the Wrapper phase.
   b. In the selected feature set, it weights the feature in accordance with the amounts of mutual information, the purpose of which is to make the initial population of the GA with a better initial point, so that the GA can use small evolution algebra and small population size to search the optimal feature subset.
2) Wrapper phase: Using the GA as the search strategy, considering the performance of the classifier and dimension of subset size as a subset evaluation criterion, search the best subset of features.

### 2.4. Features Pre-Selection Based on Mutual Information

Feature pre-selection based on mutual information, which is essentially a feature selection method based on correlation. The measurements of the correlation between the two attributes are: linear correlation coefficient and mutual information. The linear correlation coefficient of attribute *X* and *Y* is defined as:

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y}_i)}{\sqrt{\sum_i (x_i - \bar{x})^2} \sqrt{\sum_i (y_i - \bar{y})^2}} \quad (8)$$

Where $\bar{x}$ is the mean of *X*, $\bar{y}$ is the mean of *Y*, $r \in [-1,1]$]. If *X* and *Y* are linear correlated, then *r*=1 or *r*=-1. If *X* and *Y* are linearly un correlated, then r=0.

The disadvantage of the linear correlation coefficient is that it cannot capture the non-linear correlation between attributes, and can only be used to calculate the correlation between numerical attributes. In order to overcome the shortcomings of the linear correlation coefficient, mutual information is used as a measure of the correlation between attributes in this paper. For the definition and calculation of mutual information, please refer to the 2.1 section of this paper.

To sum up the above arguments, the feature pre-selection algorithm based on mutual information is as follows:

**Algorithm 1**. Feature pre-selection based on mutual information

**Input:** training data set D, original feature set $F = \{f_1, f_2, \ldots, f_L\}$, the threshold $\delta$ of *SU*.

**Output:** the selected features subset $F'$ in descending order.

**Step 1:**

initialize $F'$ to the empty set;

**Step 2:**

for *i* = 1 to *N*

According to formula (1) to (7) to calculate the $SU_i$ value between each feature $f_i$ and category label;

If ($SU_i \geq \delta$) put $f_i$ to $F'$

end for

**Step 3:**

Sort $F'$ according to the value of $SU_i$ in descending order.

On the one hand, the algorithm implements a preliminary screening of the original features set, at the same time, the sorted result of features if obtained. This sorted result will be further used in the subsequent GA.

### 2.5. Wrapper Method Based on GA

In the wrapper phase of HFS-MIGA, GA is used to search the final feature subset in the feature set which is selected from original feature set with mutual information. Detail steps are as follows:

**Step 1:** GA coding method.

Using the binary code, the length of the code is used as the number of features in the feature subset output from **algorithm 1**. If the value of a certain bit in the binary string is 1, it indicates that the corresponding feature is selected, and vice versa. The advantage of using binary coding is that crossover, mutation and other genetic operations are easy to implement.

**Step 2**: Population initialization.

GA initialization process referred to the practice of the literature [18]. The initialization of the population is guided by the feature-sorted-result obtained from the filter phase based on mutual information. GA is a random search algorithm and its performance is easily affected by the initial population, a good initial population can provide a good search starting point for GAs. The feature sorted result based on mutual information is equivalent to providing a priori knowledge of the problem for GA. The specific steps for population initialization of GA are:

a. Using of the algorithm 1 to screen and sort the features. The algorithm 1 outputs the pre-selected and sorted feature subset $F'$, $F'$ is sorted according to the value of $SU_i$ in descending order.
b. Generate the selection probabilities of each feature: set the probability to be $p_1$ for the first feature in the feature subset $F'$,, and $p_2$for the last feature in the feature subset $F'$, and generate the probabilities for the other features according to arithmetic sequence [12], where $p_1$ and $p_2$are artificially set parameters.
c. Population initialization: suppose the population has g individuals, according to the alternative probability of the feature, to generate g individuals, that is, g features subset.

**Step 3:** The design of individual fitness function.

The design of individual fitness function is a key step in GA. The individual fitness function is chosen to consider two factors: the classification accuracy and the dimension of the feature subset. So, the second phase of HFS-MIGA is a kind of Wrapper method. Fitness function should be met:

a. When the classification accuracy is higher, the function value is greater.
b. When the feature subset is smaller, the value of the function is greater.

We design the fitness function as shown in formula (9), the advantage of this function is that we can control the contribution which is contributed by the dimension of feature subset and the accuracy of classification by adjusting the parameter α.

$$f(X) = \alpha \cdot exp\big(1 - error(X)\big) + exp\left(1 - \frac{|X|}{N}\right) \quad (9)$$

error(*X*) represents the error rate of the classifier, |*X*| represents the dimension of a feature subset, *N* represents the dimension of the original feature set, α as control parameter. Typically set α>1, that is, the classification accuracy of feature subset is more important.

## 3. Results and Analysis

### 3.1. Experimental Data Set

We did two experiments. In the first experiment, we choose the popular face data set Yale [6, 7], Figure 1. show a part of face image in the database. There are 15 people faces in Yale database, and each people have 11 face images. In the experiment, all face images are cut into 32*32 pixel, and put the eyes in the same location by hand.

Figure 1. A part of face images from Yale database

In the second experiment, we selected seven data sets from the UCI machine learning repository [19] to test the performance of HFS-MIGA. Table 1 lists the names of the seven benchmark data sets, the number of features, the number of samples, and the number of categories [20, 21].

Table 1. Basic Information of Experimental Data Sets

| Dataset | Instances | Classes | Features |
|---|---|---|---|
| Anneal | 898 | 6 | 38 |
| Breast cancer | 569 | 2 | 30 |
| Dermatology | 366 | 6 | 33 |
| Ionosphere | 351 | 2 | 34 |
| Lung cancer | 32 | 2 | 56 |
| Sonar | 208 | 2 | 60 |
| Soybean-large | 307 | 19 | 35 |

### 3.2. Parameter Setting

In order to ensure that the algorithms are in a fair text environment, data sets are divided into three parts: the training set, the validation set and the test set. Training set is used to learn the selected subspace of algorithm. Validation set is to get the dimension of the best feature subspace. Test set is used to test classification rate in the best feature subspace.

In the feature pre-selection algorithm based on mutual information, we set the threshold δ=0.1 of *SU* and set the α=5 in formula (9). In order to avoid too many features is deleted in this step, resulting in subsequent GA search space is too small. If the output of the optimal feature subset is less than 1/3 of the number of original features, the first 1/3 features are retained.

The parameters of GA are set as follows: the population size was 20, the maximum number of iterations was 20, the crossover probability was 0.6, and the mutation probability was 0.033. Select operation using the proportional selection operator, and use the optimal preservation strategy. The individuals who are the most adaptive to the current population do not participate in the crossover and mutation operation; instead, it is used to replace the individual with the lowest degree of adaptation after genetic operations such as crossover and mutation. Crossover operator uses the single point crossover operator; mutation operator uses the basic bit mutation operator; the termination condition is to achieve the maximum number of iterations or continuous 5 times the optimal solution remains unchanged.

### 3.3. Experimental Results and Analysis

For the first experiment with Yale database, since it contains 11 face images of each person, so we can randomly choose (3,5,7) samples as training set, half of the samples in the remaining as test set, the other half as validation set. We independently ran algorithm 10 times to get the final average recognition error rate. Figure 2 to Figure 4 show the identification error rate in validation set corresponding to each dimensional subspace, when the number of training samples is 3, 5 or 7 in each class.

Table 2 shows the dimension of optimal subspace in various algorithms and on the classification rate in the test set. We can conclude that our algorithm on the classification error rate is lower than standard Symmetrical Uncertainty mutual information algorithm algorithm (SU) and the GA wrapper method.

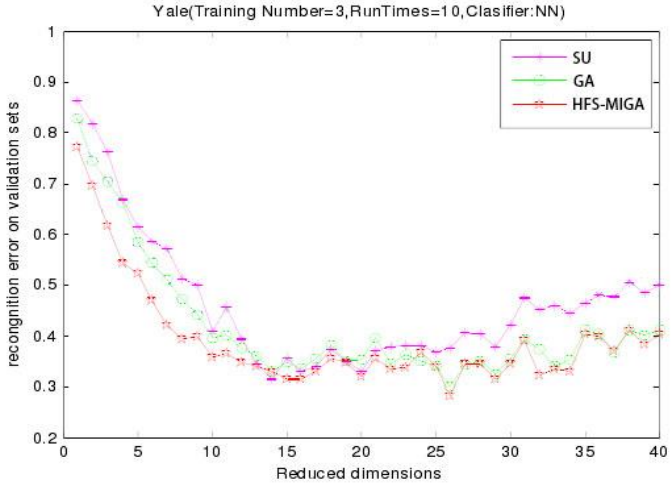


Figure 2. Recognition Error on Validation Set of Yale Data Base with 3 Training to Various Dimensions of Features

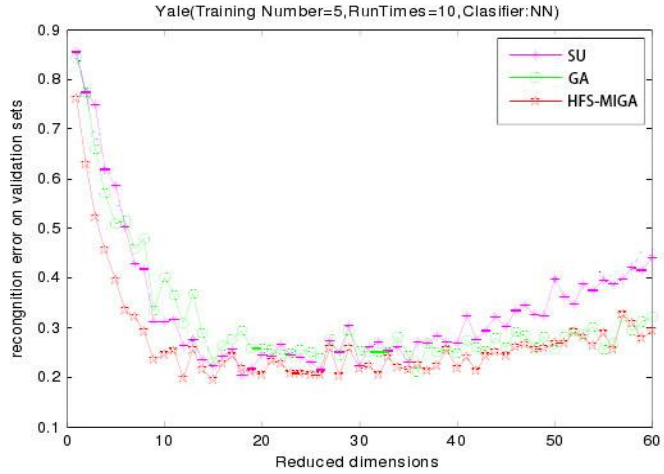


Figure 3. Recognition error on validation set of Yale data base with 5 training to various dimensions of features

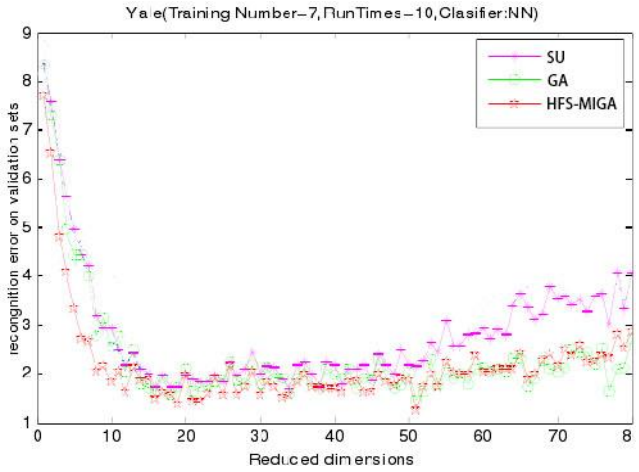


Figure 4. Recognition error on validation set of Yale data base with 7 training to various dimensions of features

Table 2. The Best Classification Error Rateof Test Set in Yale Database (Mean ±STD (Dimension of Best Feature Subspace))

| Number of Training sample | 3 | 5 | 7 |
|---|---|---|---|
| SU | 50.16±4.71 (36) | 40.44±3.79 (31) | 34.33±6.86 (23) |
| GA | 39.50±4.78 (13) | 27.56±6.39 (12) | 23.67±7.24 (14) |
| HFS-MIGA | 33.67±3.99 (26) | 20.00±5.24 (15) | 16.33±7.67 (51) |

In order to verify the effectiveness of the proposed two phases feature selection method, this paper selects a representative filter method (standard Symmetrical Uncertainty mutual information algorithm, abbreviated as SU) and the GA wrapper method, and the HFS-MIGA algorithm is compared with each other. The threshold value of the SU algorithm is set to 0.1, that is, the features of the normalized mutual information less than that of the value will be removed. In addition to without the pre-section phase based on standard mutual information, GA and HFS-MIGA algorithm adopt the same parameter setting. That is to say, GA operates directly in the data feature space of without the pre-selection phase based on standard mutual information, and selects the best individual as the result of the feature selection.

Naive Bayes (NB) is used as classifier in the experiment. For each data set in Table 1, run the SU, GA and HFS-MIGA algorithms respectively, and the size of the optimal feature subset obtained by each algorithm is recorded in Table 3.

Table 3. Feature Subset Size Obrainted from SU, GA and HFS-MIGA

| Dataset | Original features | Optimal features | | |
|---|---|---|---|---|
| | | SU | GA | HFS-MIGA |
| Anneal | 38 | 21 | 8 | 5 |
| Breast cancer | 30 | 18 | 8 | 3 |
| Dermatology | 33 | 24 | 10 | 9 |
| Ionosphere | 34 | 32 | 9 | 10 |
| Lung cancer | 56 | 12 | 25 | 4 |
| Sonar | 60 | 14 | 16 | 6 |
| Soybean-large | 35 | 21 | 17 | 13 |
| Average | 40.86 | 20.29 | 13.29 | 7.14 |

Then through the 10-fold cross validation method, these feature subsets are used in NB to classify the seven data sets, and the corresponding classification accuracy is calculated and recorded in Table 4.

Table 4. NB Classification Accuracy on Each FEture Subset (%)

| Dataset | Original features | Accuracy on optimal features | | |
|---|---|---|---|---|
| | | SU | GA | HFS-MIGA |
| Dermatology | 97.26 | 97.27 | 98.91 | 98.91 |
| Lung cancer | 84.38 | 87.50 | 90.63 | 90.63 |
| Breast cancer | 92.97 | 92.97 | 95.95 | 96.84 |
| Soybean-large | 92.18 | 92.18 | 93.81 | 92.84 |
| Ionosphere | 82.62 | 82.91 | 92.02 | 92.84 |
| Anneal | 98.44 | 98.66 | 98.78 | 98.78 |
| Sonar | 71.15 | 77.40 | 76.44 | 82.86 |
| Average | 88.43 | 89.84 | 92.36 | 93.33 |

From Table 3 and 4 we can see that the Su, GA and three algorithms in the seven data sets on the average accurate rate is higher than average accuracy using the full feature set, feature selection is indeed an effective way to improve the performance of classifier. Compared

with SU and GA algorithms, due to the feature ranking results obtained from the pre-section phase based on mutual information, in the second phase the HFS-MIGA can have a better search starting point, to obtain better feature subsets and ensure the classification efficiency of the NB classifier. In addition, the average accuracy of the HFS-MIGA algorithm on the seven data sets is higher than SU and GA, only on the soybean-large data set is lower than GA, and in the rest of the data set is not less than SU and GA. Therefore, the HFS-MIGA can greatly reduce the number of features of the data set, and achieve a better feature reduction effect when the classification accuracy is guaranteed.

## 4. Conclusion

A hybrid feature selection algorithm based on mutual information and GA is proposed in this paper. This algorithm is combination of filter and wrapper method; it is a two-phase algorithm. In the filter pre-selection phase, remove the features with lower SU value. In the wrapper phase, use the results of filter phase to guide the initialization of GA population and obtained the final optimal feature subset. The algorithm combines the advantages of the filter method and the wrapper method. The experimental results show that the algorithm is effective, which can not only guarantee the accuracy of classification, but also reduce the number of features of the data set.

## Acknowledgement

The author would like to acknowledge Huizhou University Natural Science Fund Committee for the financial support of this research. And this research is supported by the Natural Science Foundation of Huizhou University, Grant No.2016YB14

## References

[1] Lan YD, Deng H, Chen T. *Dimensionality reduction based on neighborhood preserving and marginal discriminant embedding.* Procedia Engineering. 2012 Jan 1; 29: 494-8.
[2] Chandrashekar G, Sahin F. A survey on feature selection methods. *Computers & Electrical Engineering.* 2014 Jan 31; 40(1): 16-28.
[3] Ahmad SR, Yaakub MR, Bakar AA. Detecting Relationship between Features and Sentiment Words using Hybrid of Typed Dependency Relations Layer and POS Tagging (TDR Layer POS Tags) Algorithm. *International Journal on Advanced Science, Engineering and Information Technology.* 2016 Dec 9; 6(6): 1120-6.
[4] Sun X, Liu Y, Li J, Zhu J, Chen H, Liu X. Feature evaluation and selection with cooperative game theory. *Pattern recognition*. 2012 Aug 31; 45(8): 2992-3002.
[5] Farahat AK, Ghodsi A, Kamel MS. Efficient greedy feature selection for unsupervised learning. *Knowledge and information systems*. 2013 May 1; 35(2): 285-310.
[6] Hou C, Nie F, Li X, Yi D, Wu Y. *Joint embedding learning and sparse regression: A framework for unsupervised feature selection.* IEEE Transactions on Cybernetics. 2014 Jun; 44(6): 793-804.
[7] Kumar V, Minz S. Feature Selection. SmartCR. 2014 Jun; 4(3): 211-29.
[8] Silvestre C, Cardoso MG, Figueiredo M. Feature selection for clustering categorical data with an embedded modelling approach. *Expert systems*. 2015 Jun 1; 32(3): 444-53.
[9] Gupta P, Jain S, Jain A. A Review Of Fast Clustering-Based Feature Subset Selection Algorithm. *International Journal of Technology Enhancements and Emerging Engineering* Research. 2014 Nov 25; 3(11): 86-91.
[10] Leordeanu M, Radu A, Sukthankar R. Features in concert: *Discriminative feature selection meets unsupervised clustering*. arXiv preprint arXiv:1411.7714. 2014 Nov 27.
[11] Arbain NA, Azmi MS, Ahmad SS, Nordin R, Mas' ud MZ, Lateh MA. Detection on Straight Line Problem in Triangle Geometry Features for Digit Recognition. *International Journal on Advanced Science, Engineering and Information Technology*. 2016 Dec 4; 6(6): 1019-25.
[12] Bermejo P, de la Ossa L, Gámez JA, Puerta JM. Fast wrapper feature subset selection in high-dimensional datasets by means of filter re-ranking. *Knowledge-BasedSystems*. 2012 Feb 29; 25(1): 35-44.
[13] Gan JQ, Hasan BA, Tsui CS. A filter-dominating hybrid sequential forward floating search method for feature subset selection in high-dimensional space. *International Journal of Machine Learning and Cybernetics*. 2014 Jun 1; 5(3): 413-23.

[14] Yang P, Liu W, Zhou BB, Chawla S, Zomaya AY. *Ensemble-based wrapper methods for feature selection and class imbalance learning.* InPacific-Asia Conference on Knowledge Discovery and Data Mining 2013 Apr 14 (pp. 544-555). Springer Berlin Heidelberg.
[15] Schneider TD. *Information theory primer with an appendix on logarithms.* InNational Cancer Institute 2007.
[16] Yu L, Liu H. Feature selection for high-dimensional data: A fast correlation-based filter solution. *InICML 2003 Aug 21* (Vol. 3, pp. 856-863).
[17] Fogel LJ, Owens AJ, Walsh MJ. Adaptation in natural and artificial systems.
[18] Richard OD, Peter EH, David GS. *Pattern classification. A Wiley-Interscience*. 2001:373-8.
[19] Frank A, Asuncion A. UCI machine learning repository.
[20] Zheng W, Feng G. Feature Selection Method Based on Improved Document Frequency. *TELKOMNIKA Telecommunication Computing Electronics and Control.* 2014 Dec 1; 12(4): 905-10.
[21] Zhu SX, Hu B. Hybrid feature selection based on improved GA for the intrusion detection system. *Indonesian Journal of Electrical Engineering and Computer Science*. 2013 Apr 1; 11(4): 1725-30.