**American Journal of Software Engineering and Applications**
2014; 3(6): 68-73
Published online December 05, 2014 (http://www.sciencepublishinggroup.com/j/ajsea)
doi: 10.11648/j.ajsea.20140306.11
ISSN: 2327-2473 (Print); ISSN: 2327-249X (Online)

# The analysis of GCFS algorithm in medical data processing and mining

**Xiao Yu Chen, Bo Liu*, Zhe Feng Zhang, Xin Xia**

Department of Information Centre, East Hospital, Tongji University, School of Medicine, Shanghai, China

**Email address:**
alexander191@sina.com (Xiao Yu Chen), liubonew@126.com (Bo Liu), zhang_zf26@126.com (Zhe Feng Zhang),
xinye_000@163.com (Xin Xia)

**To cite this article:**
Xiao Yu Chen, Bo Liu, Zhe Feng Zhang, Xin Xia. The Analysis of GCFS Algorithm in Medical Data Processing and Mining. *American Journal of Software Engineering and Applications.* Vol. 3, No. 6, 2014, pp. 68-73. doi: 10.11648/j.ajsea.20140306.11

**Abstract:** Feature selection plays a significant part in medical data processing and mining, it can reduce the dimensionalities of datasets and enhance the performance of the classifiers, and it is also helpful to clinical decision support to a great extent. At present, the clinical decision support is more performed by physicians subjectively based on clinical knowledge, which may hinder the diagnosis and treatment. This paper mainly outlines the performance of GCFS (Genetic Correlation-based Feature Selection) algorithm in the processing and mining procedure of medical data, and medical UCI datasets are employed as the studied materials for proving the improvement of feature selection in data classification. Compared with the algorithms of CFS and GA (Genetic Algorithm), ensemble learning methods are employed as the testing classifiers, and the results show GCFS algorithm almost improves the performances of the testing classifiers better than CFS and GA.

**Keywords:** Feature Selection, GCFS, Ensemble learning

## 1. Introduction

The applications of computer and information technology have made a new highlight research direction with a rapid development in medicine. In clinics, medical diagnosis is considered as a classification problem: a case represents a patient's information, condition features are the patients' data and the category is the diagnosis, so it is essential to build classification model which can predict the uncategorized cases. Medical datasets are inevitable to contain irrelevant and noisy features, the selection of appropriate subset of the available features can produce compact and interpretable results for modeling the data adequately, and it can improve the classification accuracy in medical region [1].

Some research efforts have been devoted to the applications of data mining techniques for discovering useful medical knowledge and rules; such as Wang et al. [2] proposed a DFP-growth (Database Frequent Pattern) feature selection algorithm for the classification of children pneumonia cases; H. M. Yan proposed a real-coded genetic method to select critical features essential to the heart diseases diagnosis, and the critical features and their clinic meanings are in sound agreement with those used by the physicians in making their clinic decisions [3]. R. E. Abdel-Aal used the group method of data handling (GMDH) to reduce the data dimensionalities for the breast cancer and heart disease, and it also lead to the improvements in the overall classification performance [4].

In this paper, the feature selection algorithm of GCFS is adopted in the mining procedure of medical data, and it is compared with CFS and GA based on the ensemble learning classifiers of Bagging and Boosting methods for demonstrating its suitable application and better performance. This paper is organized as follows. Section 2 describes the studied materials, and the data processing and mining procedure and methods are shown in Section 3. The results and analysis about GCFS are given in Section 4 and 5, and Section 6 summarizes this paper and gives the conclusions.

## 2. Materials

In this paper, there are four UCI medical datasets employed as the studied materials, they are downloaded at http://archive.ics.uci.edu/ml/datasets.html, including: diabetes, Breast cancer, Hepatitis survival, and CTGs (cardiotocograms).

Table 1 shows the description about the studied datasets. In the datasets, the missing feature values are replaced by mean values of the corresponding features, and missing values of all the features are less than 1%.

***Table 1.*** *Description of the studied medical datasets*

| Dataset Name | Categories | Features | Instances | Data type | Category Instruction |
|---|---|---|---|---|---|
| Diabetes | 2 | 8 | 768 | Continuous | negative=0, positive=1 |
| Breast cancer | 2 | 9 | 699 | Nominal | benign=2, malignant =4 |
| Hepatitis survival | 2 | 19 | 155 | Mixed | live=1, die=2 |
| CTGs | 3 | 21 | 2126 | Continuous | normal=1, suspect=2, pathologic=3 |

# 3. Methods

In this paper, the data processing course includes GCFS algorithm for feature selection, ensemble learning, and C4.5 decision tree for classification, and Figure 1 gives the framework of data process for instruction.

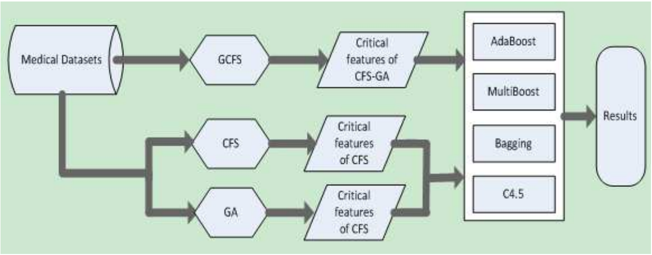


***Figure 1.*** *Framework of data processing and classification*

## *3.1. GCFS algorithm*

GCFS algorithm is the attribute selection part in the data process. CFS (Correlation-based Feature Selection) is a classical filtered algorithm of attribute selection; in this algorithm, the heuristic evaluation for a single feature corresponding to each category label is used to obtain the final feature subset, and the assessment method of CFS is as follows:

$$Ms = \frac{k\overline{r_{cf}}}{\sqrt{k+k(k-1)+\overline{r_{ff}}}} \tag{1}$$

In (1), Ms is the evaluation for an attribute subset S including k attribute items, $\overline{r}_{cf}$ is the mean correlation degree between attributes and the category label, and $\overline{r}_{ff}$ is the mean correlation degree among attributes. And the evaluation of CFS is a method of correlation based on attribute subsets. A bigger $\overline{r}_{cf}$ or smaller $\overline{r}_{ff}$ in acquired subsets by the method produce a higher evaluation value, and in CFS, the correlation degree among attributes is calculated by information gain, and the formula of information gain is shown below. Y is the category attribute, y is any possible value of Y, the entropy of Y is shown in (2), and for an attribute X, entropy of category attribute Y under the condition of X is in (3).

$$H(Y) = -\sum_{y\in Y} p(y)\log_2(p(y)) \tag{2}$$

$$H(Y|X) = -\sum_{x\in X} p(x)\sum_{y\in Y} p(y|x)\log_2(p(y|x)) \tag{3}$$

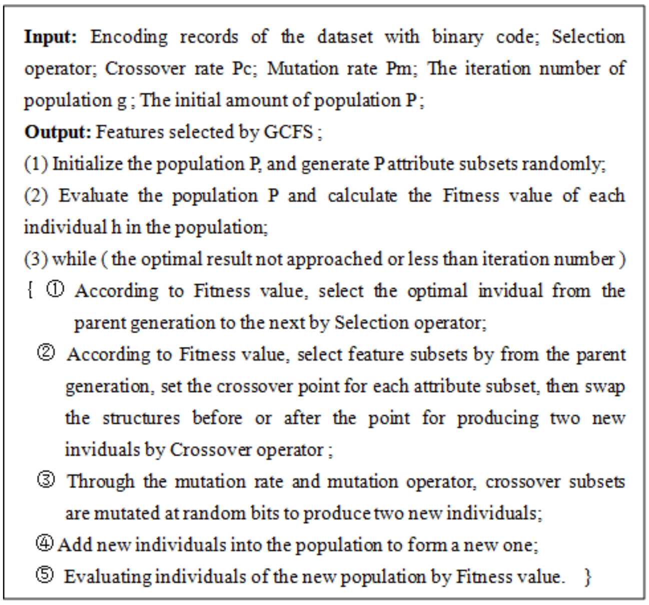
The difference of H (Y) － H (Y|X) (i.e. the entropy reduction of attribute Y) can reflect the information amount provided by attribute X to attribute Y, and a bigger difference means a higher correlation degree between X and Y. Information gain is a symmetrical evaluation method; it tends to select the attributes with more values. Therefore, it is necessary to normalize information gain to [0, 1] for keeping equivalent comparison effect among attributes, and (4), below, shows the calculating formula.

$$U_{XY} = 2.0\times\frac{H(Y)-H(Y|X)}{H(Y)+H(X)} \tag{4}$$

As a filtering algorithm, CFS evaluates the correlation between attributes and category label, and the redundancy degree among attributes [5]. Although the algorithm performs well in dimension reduction, it cannot approach a global optimum result. The Genetic algorithm (GA) is a wrapping algorithm in dimension reduction for its global search capability [6-8]. In this paper, CFS and GA are combined to make the GCFS algorithm, and this algorithm evaluates new individuals of GA through the correlation degree in CFS as the fitness function of GA. The design of GCFS algorithm mainly includes four parts: coding scheme, selection operator, crossover operator and mutation operator.

In the coding scheme, each entity is encoded with classical binary code. The method of roulette wheel is employed for selection operator. For the crossover operator, single- point crossover is used to produce new individuals by swapping the cross point part through the crossover points. And basic bit mutation is used in binary encoding for the mutation operator, from 0 to 1, or from 1 to 0.

In the selection of the crossover rate and mutation rate, for producing more new individuals and avoiding causing too much damage to the better attribute subset, the crossover rate range is from 0.40 to 0.99 and the mutation rate is from 0.0001 to 0.1 commonly. The description of GCFS algorithm is shown in Figure 2.



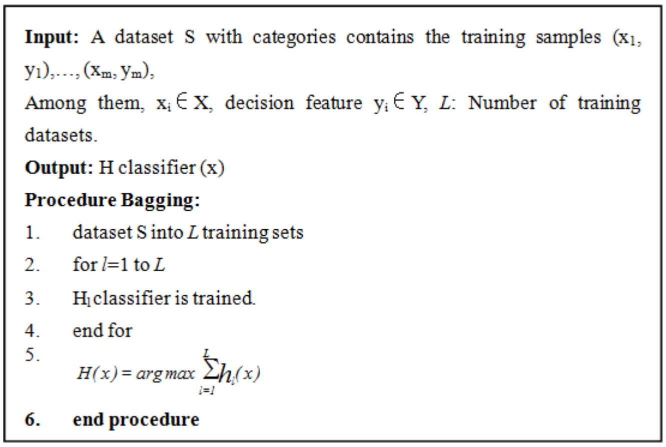
```
Input: Encoding records of the dataset with binary code; Selection operator; Crossover rate Pc; Mutation rate Pm; The iteration number of population g ; The initial amount of population P ;
Output: Features selected by GCFS ;
(1) Initialize the population P, and generate P attribute subsets randomly;
(2) Evaluate the population P and calculate the Fitness value of each individual h in the population;
(3) while ( the optimal result not approached or less than iteration number )
{ ① According to Fitness value, select the optimal invidual from the parent generation to the next by Selection operator;
  ② According to Fitness value, select feature subsets by from the parent generation, set the crossover point for each attribute subset, then swap the structures before or after the point for producing two new inviduals by Crossover operator ;
  ③ Through the mutation rate and mutation operator, crossover subsets are mutated at random bits to produce two new individuals;
  ④ Add new individuals into the population to form a new one;
  ⑤ Evaluating individuals of the new population by Fitness value.  }
```

***Figure 2.*** *Description of GCFS algorithm*

### *3.2. Ensemble Methods*

Ensemble methods have become a mainstay in the machine learning and data mining literature. They are designed based on “No classifier is perfect” as the guideline, and combined the performance of many weak classifiers to produce a powerful committee. Boosting and bagging are popular choices.

The main idea of Boosting algorithm [9] is the learning enhancement based on the misclassified samples. At first, each sample is endowed with the same weight, then the first basic classifier is employed to classify the samples and test the training dataset based on the weights. For the misclassified samples, their weights will be upgraded, and the second classifier will be trained on the dataset with modified weights by iteration until an optimal classifier can be obtained. Based on the idea, various boosting algorithms are proposed for different problems, such as AdaBoost [10], and MultiBoost [11].



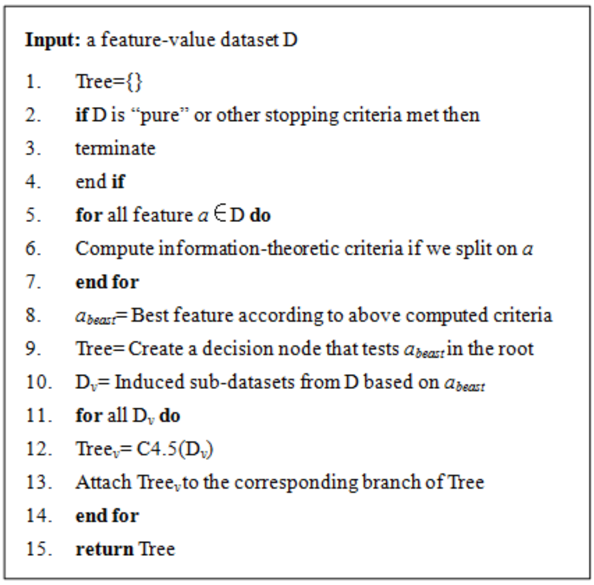
```
Input: A dataset S with categories contains the training samples $(x_1, y_1),\ldots,(x_m, y_m)$,
Among them, $x_i \in X$, decision feature $y_i \in Y$, $L$: Number of training datasets.
Output: H classifier (x)
Procedure Bagging:
    dataset S into $L$ training sets
    for $l$=1 to $L$
        $H_l$ classifier is trained.
    end for
    $H(x) = \arg\max \sum_{l=1}^{L} h_l(x)$
end procedure
```

***Figure 3.*** *Bagging Algorithm*

Bagging (Bootstrap Aggregating) is a method based on resampling technique [12]. In Bagging method, a weak learning algorithm and a training set $((x_1, y_1), \ldots ,(x_m, y_m))$ are given, the algorithm generates a number of training sets including some samples by random from the initial training set, and the training sets and the initial training sets are nearly in the same size. Samples of the initial training set in a round may appear or not, and the learning algorithm classifier is trained on each training set and obtains a predicted sequence $g_1, g_2, \ldots ,g_t$, and the final function G can be predicted by voting. And the description of Bagging algorithm is shown in Figure 3.

### *3.3. C4.5*

C4.5 [13] is a suite of algorithms for classification problems in machine learning and data mining. It aims at supervised learning: Given an attribute-valued dataset where instances are described by collections of attributes and belong to one of a set of mutually exclusive classes, and C4.5 learns a mapping from attribute values to classes that can be applied to classify new, unseen instances. The generic description of how C4.5 works is given in Figure 4. All tree induction methods begin with a root node that represents the entire, given dataset and recursively split the data into smaller subsets by testing for a given attribute at each node. The sub trees denote the partitions of the original dataset that satisfy specified attribute value tests. This process typically continues until the subsets are “pure,” that is, all instances in the subset fall in the same class, at which time the tree growing is terminated. Considering the characteristics of the datasets, C4.5 decision tree is employed and it is fit for continuous or nominal features in datasets.



```
Input: a feature-value dataset D
Tree={}
if D is “pure” or other stopping criteria met then
terminate
end if
for all feature $a \in$ D do
Compute information-theoretic criteria if we split on $a$
end for
$a_{beast}$= Best feature according to above computed criteria
Tree= Create a decision node that tests $a_{beast}$ in the root
$D_v$= Induced sub-datasets from D based on $a_{beast}$
for all $D_v$ do
Tree$_v$= C4.5($D_v$)
Attach Tree$_v$ to the corresponding branch of Tree
end for
return Tree
```

***Figure 4.*** *Description of C4.5 Algorithm*

## 4. Results

### *4.1 Feature Selection of GCFS*

Feature reduction is an approach for improving the performances for classifiers, the GCFS algorithm parameters are defaulted as follows: the population size and the number of generations is 20, the probabilities of crossover and mutation are 0.6 and 0.033 respectively. In Table 2, we can see that GCFS can achieve the purpose of feature reduction on the medical datasets through ten-fold cross validation. The feature reduction rate of GCFS is up to 26.31%.

***Table 2.*** *The feature selection rate of GCFS*

| Datasets | Full Features | GCFS |
|---|---|---|
| Diabetes | 8 | 4 |
| Breast cancer | 9 | 9 |
| Hepatitis survival | 19 | 13 |
| CTGs | 21 | 16 |
| Ave. feature number | 14.25 | 10.5 |
| Feature reduction rate (%) | 0 | 26.31% |

### *4.2. Performances of GCFS on the Testing Classifiers*

To evaluate the performances of GCFS on the testing classifiers, two criteria of ACC (classification accuracy) and AUC (area under ROC curve) are used within the framework.

- ACC is calculated as the percentage of the correctly classified testing samples over all the test samples.
- AUC is a relative evaluation standard, and it has been recently proposed as an alternative single-number measure for evaluating the predictive ability of learning algorithms [18].

Through ten-fold cross validation, ACC (%) and AUC of the three types of testing classifiers (Bagging, Boosting and C4.5 decision tree) on the UCI medical datasets filtered from GCFS are listed in Table 3.and Table 4. For meeting the homogeneity of the testing classifiers, C4.5 is also employed as the basic classifiers of testing ensemble learning methods of AdaBoostM1, MultiBoostAB, and Bagging. From Table 3 and Table 4, we can see GCFS performs well in medical data classification of the testing classifiers.

The testing classifiers are running based on the hardware environment of an Intel Core 2 Duo CPU 2.4GHz and a Memory of 2G. Figure 5 shows the time-running status of the testing classifiers on the medical datasets filtered from GCFS, and we can see that the running times of the testing classifiers are in fast speeds and acceptable.

***Table 3.*** *ACC(%)of the testing classifiers on GCFS*

| | AdaBoostM1 | MultiBoostAB | Bagging | C4.5 |
|---|---|---|---|---|
| Diabetes | 72.47 | 74.76 | 74.49 | 71.79 |
| Breast cancer | 95.85 | 95.85 | 95.42 | 94.85 |
| Hepatitis survival | 83.23 | 81.29 | 81.29 | 80 |
| CTGs | 94.78 | 94.83 | 94.31 | 93.09 |

***Table 4.*** *AUC of the testing classifiers on GCFS*

| | AdaBoostM1 | MultiBoostAB | Bagging | C4.5 |
|---|---|---|---|---|
| Diabetes | 0.761 | 0.789 | 0.812 | 0.745 |
| Breast cancer | 0.983 | 0.990 | 0.983 | 0.968 |
| Hepatitis survival | 0.780 | 0.806 | 0.787 | 0.661 |
| CTGs | 0.982 | 0.983 | 0.976 | 0.924 |

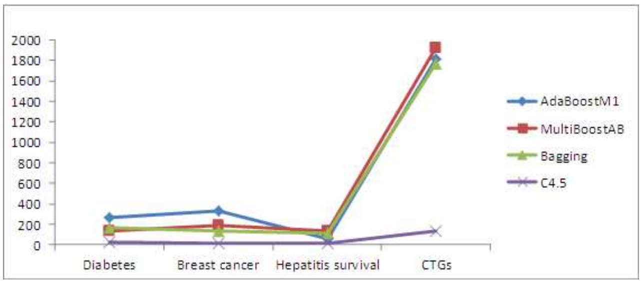


***Figure 5.*** *Running times of the testing classifiers on the datasets from GCFS (ms)*

### *4.3.Comparisons with CFS and GA*

GCFS in feature selection is compared with the algorithms of CFS and GA on through ten-fold cross validation. Table 5 gives a comparison of the feature selection approaches on the UCI medical datasets, the reduction rates of CFS and GA are 42.11% and 7.02% in feature selection, and it shows that the GCFS makes a higher rate of feature reduction than GA, but it is lower than that of CFS.

From Table 6 to Table 9, we can see the classifiers of AdaBoostM1, MultiBoostAB, Bagging and C4.5 generally perform well based on the feature selection methods of CFS and GA. Figure 6 and Figure 7 show the running times of the testing classifiers on CFS and GA, and we can see the classifiers based on CFS consume shorter running times than based on GA.

***Table 5.*** *The comparison of feature selection rates with CFS and GA*

| Datasets | GCFS | CFS | GA |
|---|---|---|---|
| Diabetes | 4 | 4 | 7 |
| Breast cancer | 9 | 9 | 9 |
| Hepatitis survival | 13 | 12 | 17 |
| CTGs | 16 | 8 | 20 |
| Ave. feature number | 10.5 | 8.25 | 13.25 |
| Feature reduction rate (%) | 26.32 | 42.11 | 7.02 |

***Table 6.*** *ACC(%) of the testing classifiers on CFS*

| | AdaBoostM1 | MultiBoostAB | Bagging | C4.5 |
|---|---|---|---|---|
| Diabetes | 72.47 | 74.76 | 74.49 | 71.79 |
| Breast cancer | 95.85 | 95.85 | 95.42 | 94.85 |
| Hepatitis survival | 83.87 | 85.81 | 81.94 | 80 |
| CTGs | 93.79 | 94.03 | 93.41 | 93.04 |

***Table 7.*** *AUC of the testing classifiers on CFS*

| | AdaBoostM1 | MultiBoostAB | Bagging | C4.5 |
|---|---|---|---|---|
| Diabetes | 0.761 | 0.789 | 0.812 | 0.745 |
| Breast cancer | 0.983 | 0.990 | 0.983 | 0.968 |
| Hepatitis survival | 0.827 | 0.817 | 0.810 | 0.651 |
| CTGs | 0.97 | 0.973 | 0.968 | 0.916 |

***Table 8.*** *ACC(%) of the testing classifiers on GA*

| | AdaBoostM1 | MultiBoostAB | Bagging | C4.5 |
|---|---|---|---|---|
| Diabetes | 72.20 | 74.36 | 74.36 | 73.14 |
| Breast cancer | 95.85 | 95.85 | 95.42 | 94.85 |
| Hepatitis survival | 84.52 | 83.23 | 82.58 | 81.94 |
| CTGs | 94.40 | 94.26 | 93.89 | 92.89 |

***Table 9.*** *AUC of the testing classifiers on GA*

| | AdaBoostM1 | MultiBoostAB | Bagging | C4.5 |
|---|---|---|---|---|
| Diabetes | 0.762 | 0.788 | 0.813 | 0.754 |
| Breast cancer | 0.983 | 0.990 | 0.983 | 0.968 |
| Hepatitis survival | 0.8 | 0.799 | 0.799 | 0.693 |
| CTGs | 0.978 | 0.982 | 0.975 | 0.914 |

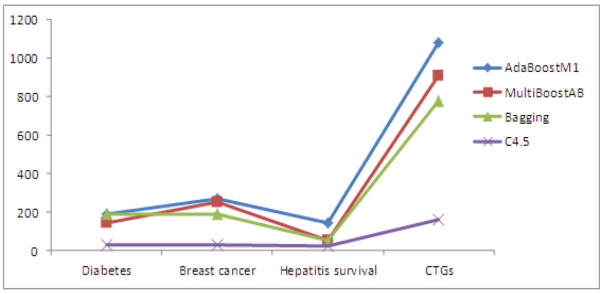


***Figure 6.*** *Running times of the testing classifiers on the datasets from CFS (ms)*

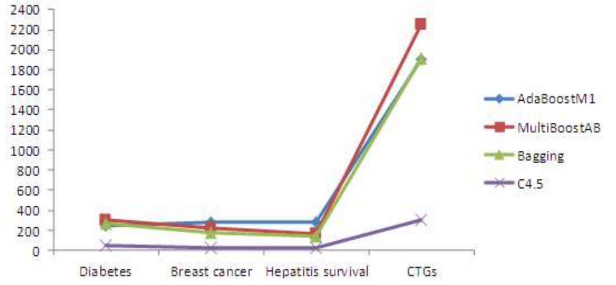


***Figure 7.*** *Running times of the testing classifiers on the datasets from GA (ms)*

## 5. Discussion

For analyzing the performances of the testing classifiers on the feature selection methods, ACC, AUC and running time of the testing classifiers are employed within the course of data processing, and Figure 8 to Figure 10 give the related results. In Figure 8 and Figure 9, we can see the average ACC values of GCFS are almost the highest on AdaBoostM1 and Bagging, the average ACC values of CFS and GA are higher than those of GCFS on MultiBoostAB and C4.5; and GCFS performs the highest average AUC values on the boosting classifiers except C4.5. Compared with the better performances in classification, GCFS shows a shorter running time than GA, but a longer time than CFS.

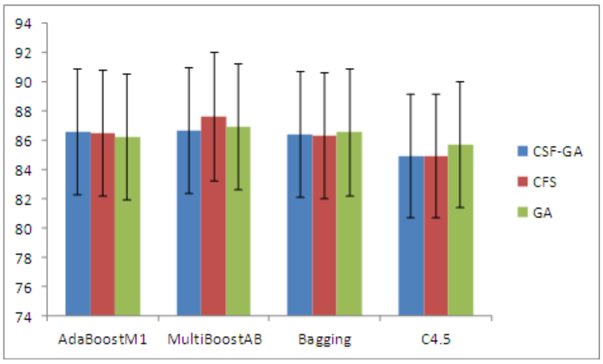


***Figure 8.*** *Ave. ACC(%) of three feature selection methods on the testing classifiers*

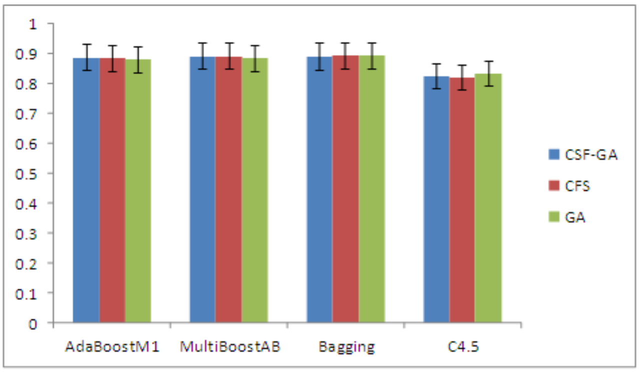


***Figure 9.*** *Ave. AUC of three feature selection methods on the testing classifiers*

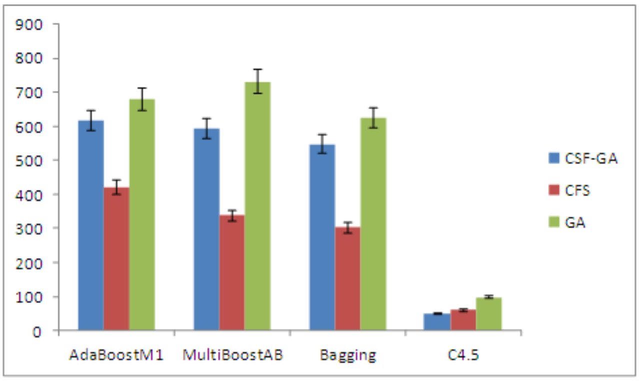


***Figure 10.*** *Ave. running times of three feature selection methods on the testing classifiers*

## 6. Conclusions

Compared with the algorithms of CFS and GA, GCFS is analyzed based on the ensemble learning methods (AdaBoostM1, MultiBoostAB and Bagging) of C4.5 in medical data classification and running times. Obviously, GCFS performs a medium reduction rate between GA and CFS in feature selection, it performs well in classification, especially a better average ACC (%) than CFS and GA on AdaBoostM1; it almost makes the highest average AUC values on the ensemble learning classifiers.; and GCFS performs a medium running time between CFS and GA on the testing ensemble learning classifiers.

## References

[1] G. Eason, B. Noble, and I. N. Sneddon, “On certain integrals of Lipschitz-Hankel type involving products of Bessel functions,” Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955. (references)

[2] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[3] I. S. Jacobs and C. P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[4] K. Elissa, “Title of paper if known,” unpublished.

[5] R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.

[6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[7] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 198.

[8] I. Skrypnik, V. Terziyan, S. Puuronen and A. Tsymbal: Proceedings of the 12th IEEE Symposium on Computer-Based Medical Systems. 1999, p. 53–58.

[9] B. Wang, M. Zhang, B. Zhang and W. Wei: Proceedings of the 7th International Conference on Parallel and Distributed Computing, Applications and Technologies. 2006, p. 128–131.

[10] H. M. Yan, J. Zheng, Y. T. Jiang, C. L. Peng, S. Z. Xiao, “Selecting critical clinical features for heart diseases diagnosis with a real-coded genetic algorithm”, Applied soft computing, no.8, (2008), pp. 1105-1111.

[11] R. E. Abdel-Aal, “GMDH-based feature ranking and selection for improved classification of medical data”, Journal of Biomedical Informatics, vol. 38, no.6, (2005), pp. 456-468.

[12] M. A. Hall, Correlation based feature selection for machine learning [D]. Hamilton, New Zealand：University of Waikato, 1999: 51-69.

[13] B. Zheng, Y. X. Jin. “The analysis of marine human error causes based on attribute reduction”, Journal of Shanghai Marine University, vol. 31, no. 1, pp. 92-93, 2010.

[14] J. T. Ren, J. H. Sun, H. Y. Huang, et al. “A feature selection method based on information gain and genetic algorithm”. Computer science, vol. 33, no. 10, pp. 194, 2006.

[15] S. C. Song, H. Pang, X. J. Ding. “The application research of GA-SVM algorithm in text classification”. Computer simulation, vol. 28, no. 1, pp. 223-225, 2011.

[16] R. E. Schapire, “The strength of weak learn ability”, Machine learning, vol. 5, no.2, (1990), pp. 197-227.

[17] Y. Freund, “Boosting a weak algorithm by majority”, Information and computation, vol.121, no.2, (1995), pp. 256-285.

[18] G. I. Webb, “MultiBoosting: A technique for combining boosting and wagging” , Machine Learning, vol. 40, no.1, (2000), pp. 159-196.

[19] J. R. Quinlan, C4.5: Programs for machine learning, Morgan Kaufmann Publishers, San Francisco, 1993.

[20] L. Breiman. Bagging predictors. Machine learning. 1996(24)：123-140.