International Journal of Database Theory and Application
Vol.9, No.11 (2016), pp.61-70
http://dx.doi.org/10.14257/ijdta.2016.9.11.06

CHI Statistical Text Feature Selection Method Based on Information Entropy Optimization

Guohua Wu[1], Sen Li[2], Lin Han[3] and Mengmeng Zhao[4]

[1, 2, 3, 4]*School of Computer Science and Technology*
Hangzhou Dianzi University, Hangzhou, China
WuGuoHua_hdu@163.com, LiSen_hdu@163.com, HL_hdu@163.com,
ZhaoMengMeng_hdu@163.com

Abstract

CHI statistical text feature selection method based on information entropy optimization is presented in this paper. In the text categorization process of feature selection, considering the results of effect of the distribution within categories and among categories, we introduce the frequency of features information entropy among categories, the information entropy within categories, information within category to optimize the CHI statistical methods. The experimental results show that the classification accuracy of the optimized CHI method is significantly higher than that the traditional CHI statistical methods.

Keywords*: feature selection, CHI statistics, information entropy optimization, text categorization*

1. Introduction

With the advent of computer technology and Internet technology, information on the Internet is exploding. Automatic text categorization has become an effective mean of obtaining knowledge and information, and widely used in the classification of mail, information retrieval, web pages detection, digital library *etc.*. However, the document represents a sparse vector and high- dimensional vector space are not only increase the time and space overhead of text automatic categorization, but also effect the text classification accuracy [1]. Therefore, feature selection is very important in the process of automatic text classification. The text feature selection method [2, 3] in common use are Document Frequency, Mutual Information, Information Gain, CHI Statistics, Expected Cross Entropy, Weight of Evidence Text *etc.* The above methods are not depend on specific classifier, but by providing an evaluation function to compare the score of the feature item, in order to select the important feature item that the feature item of lower score will be filter. Yang's *etc.*. [4, 5] results show that in the common used feature selection methods, CHI Statistics with its low time complexity and convenience, becomes one of the effective algorithm for feature selection in text categorization.

CHI statistics is less consider the feature item distribution between in different categories (short for among categories) and in the same category (short for within categories), in recent years, some scholars have made some improvements for the lack of CHI Statistics algorithm, Luigi Galavotti *etc.* [6] by analyzing the difference of the feature item in the classification, introduced the concept of positive and negative correlation between feature item and categories, and proposed a new correlation coefficient method to optimize the CHI Statistics, and the performance of CHI Statistics has been improved. In order to improve the performance of text classification, Li *etc.* [7] analysis the differences between categories of text classification, and improved the performance of text classification. Wang Guang [8] *etc.* propose a combination of CHI Statistics and

ISSN: 2005-4270 IJDTA

Information Gain feature selection algorithm to improve the lack of CHI Statistics, which based on the traditional feature selection means of CHI Statistics and Information Gain.

In order to better reflect the characteristic of the representative feature item within concentration distribution between different categories and distribution of dispersion within the same category, this paper on the basis of the existing CHI Statistics and combine the information entropy of feature frequency between different categories, Information Entropy of feature frequency in the same category and the feature frequency within the same category and other factors, proposes an CHI Statistics feature select algorithm based on information entropy optimization.

2. CHI Statistics Algorithm Based on Information Entropy Optimization

The CHI Statistics algorithm based on optimized information entropy is proposed in this paper introduce the effect of feature frequency on information entropy calculation, information entropy among categories aim at to measure the concentrate distribution between different categories of the representative feature item, information entropy within categories with the purpose of to measure the distribution of dispersion within the same category of the representative feature item, so that we can achieve an optimization CHI statistics algorithm.

In order to better explain the distribution of the feature item, we can process feature frequency within information entropy. Suppose that X is the distribution of feature frequency in the feature collection $x_1, x_2, \cdots, x_n$, $P(x_i) = p_i (0 \le p_i \le 1)$ is the probability of feature item x_i , and $p_1 + p_2 + \cdots + p_n = 1$.

The definition of $IE(X)$ [9] as follows:

$$IE(X) = -\sum_{i=1}^{n} P(x_i) \log_2 P(x_i)$$

$$= -\sum_{i=1}^{n} p_i \log_2 p_i \tag{1}$$

It can be seen that, the smaller the variance of the feature item probability distribution, the bigger the information entropy, the bigger the variance of the feature item probability distribution, the smaller the information entropy, according to this characteristic, this paper utilizes information entropy to represent the feature item distribution in automatic text classification.

2.1. Thought of Information Entropy among Categories

Information entropy among categories can be described as the probability distributions of feature item frequency in different categories in text set, if there is a feature item focused distribution in a or some categories, and less likely to appear in the other categories, the feature item will be able to properly distinguish between the category and other categories, and it contains more category information, and the Information entropy among categories is smaller. On the contrary, if a feature item is widely distributed in most categories, even though it's feature frequency is high, it does not distinguish between the categories, and contribute little to text classification, so that the feature item doesn't have to opt-in feature subset, and it's information entropy among categories is big. Information entropy among categories computing formula as follows:

$$IE_{out}(t_k) = -\sum_{j=1}^{n} P(c_j(t_k))log_2 P(c_j(t_k)) = -\sum_{i=1}^{n} \frac{f_j(t_k)}{\sum_{j=1}^{n} f_j(t_k)} log_2 \frac{f_j(t_k)}{\sum_{j=1}^{n} f_j(t_k)} \quad (2)$$

In the algorithm of information entropy among categories *j* represents the current category of the text, *n* indicates the number of text category, $f_j(t_k)$ represents the number of feature item t_k in the category c_j.

It can be seen that the larger the $IE_{\text{out}}(t_k)$, the more evenly to feature item t_k distribution between different categories, the weaker the representative of the category. The smaller the $IE_{\text{out}}(t_k)$, the more concentrated the feature item t_k distribution between different categories, the stronger the representative of the category.

2.2. Thought of Information Entropy within a Category

Information entropy within categories can be described as the probability distributions of feature item frequency within a category. If a feature item uniform distribution in a category, the feature item can be better represent the category information, and it can effectively distinguish between other categories, and the bigger the information entropy within categories. On the contrary, if a feature item only appears once in a while in a category, it means that the feature is nonuniform distribution, so that the feature item can be not better represent the category information, and the smaller the information entropy within a category. It can be obtain from the above analysis that at the time of feature item selection, in order to improve the efficiency of the classification, we should choose the feature item which reserved and feature frequency uniform distributions in different texts. Information within categories computing formula as follows:

$$IE_{\text{in}}(t_k, c_j) = -\sum_{i=1}^{m} P(t_k)\log_2 P(t_k) = -\sum_{i=1}^{m} \frac{f_{ji}(t_k)}{f_j(t_k)} \log_2 \frac{f_{ji}(t_k)}{f_j(t_k)} \quad (3)$$

In the algorithm of information entropy within categories *i* denote the ith text of category c_j, *m* denote the total number of category t_k. $f_{ji}(t_k)$ denote the total number of feature item t_k in the ith text of category c_j.

It can be seen that the larger the $IE_{in}(t_k, c_j)$, the more evenly of feature item t_k distribution within a category, and feature item t_k has better representative ability of category c_j. On the contrary, the smaller the $IE_{in}(t_k, c_j)$, feature item t_k concentrated distribution in the category, the weaker the represent of the category.

2.3. Feature Frequency Information within a Category

CHI Statistical algorithm is based on the number of documents to statistic, as a basis of the feature frequency of $t_i, i = 1,2,\cdots,m$ within the category $C_j, j = 1,2,\cdots,r$, and

use the count of document as the basic unit to calculate, but it doesn't consider the factor of feature item frequency of t_i.

Consider the following situation, we assume that feature item t_m t_k are appear in the most document d_{ji} which belong to category C_j and fewer appear in other categories, then both of t_m t_k are the important feature item to represent the information of category C_j. Through CHI Statistics formula we can achieve that the values should be close of t_m t_k within the category C_j. Setting that $n_{ji}(t_k), i = 1,2,\cdots, n_{c_j}$ is the feature item frequency of t_k within the text d_{ji}. If the feature item frequency of t_k is far outweigh the feature frequency of t_m in each text, that is

$$\sum_{i=1}^{n_{C_j}} n_{ij}(t_k) >> \sum_{i=1}^{n_{C_j}} n_{ij}(t_m) \tag{4}$$

The formula shows that feature item t_k has a great representative capacity of the document to category C_j than feature item t_m, but traditional CHI Statistics algorithm does not reflect this difference. The frequency of text which contains feature item t_k within the category C_j, and the frequency of feature item t_k in each document are important two factors for text classification, but traditional CHI Statistics feature item selection algorithm does not involve the second factor.

Generally, if feature item t_k appears in one or a few categories and it's feature item frequency is large, then t_k has a great representative of this category, we should endow feature t_k a larger weighting coefficient. Denote $tf_{ji}(t_k)$ is the frequency of feature item t_k in the text d_i within the category C_j, thus, we can get:

$$\lambda_{jk} = \sum_{i=1}^{n_{c_j}} \frac{tf_{ji}(t_k) - \min_{1\le q\le m}(tf_{ji}(t_q))}{\max_{1\le q\le m}(tf_{ji}(t_q)) - \min_{1\le q\le m}(tf_{ji}(t_q))} \tag{5}$$

When category renders oblique, and the text number of different categories also has difference, the formula above will be normalized, get the following coefficient:

$$\lambda_k = \sum_{j=1}^{r} \frac{\lambda_{jk}}{\sqrt{\sum_{k=1}^{m} \lambda_{jk}^{\ 2}}} \tag{6}$$

Obviously, the feature item t_k with a higher frequency corresponding the higher weighting coefficient λ_k, it also explains that the more frequent the feature item appears in the category, the stronger the classification abilities.

2.4. CHI Statistical Algorithm is Based on Information Entropy

The main idea of CHI Statistics algorithm based on information entropy optimization is utilize the feature item frequency which contained in the information entropy to optimize CHI Statistics algorithm. As we know, CHI Statistics algorithm is to measure the relevance of feature item and document, suppose that feature item t_k and category *c* are

fit the χ^2 distribution which fit the first degree of freedom. The value [10] of CHI statistics within the category C_j feature item t_k can be calculated as:

$$\chi^2(t_k, c_j) = \frac{N(AD - BC)^2}{(A+C)(B+D)(A+B)(C+D)} \tag{7}$$

In the formula, A represents the document frequency which contains feature item t_k and belong to category C_j, B represents the document frequency which contains feature item t_k but not belong to category C_j, C represents the document frequency which belong to category C_j but not contains feature item t_k, D represents the document frequency which not belong to category C_j or not contains feature item t_k, N represents the total number of documents in the corpus.

When feature selection, CHI Statistics algorithm is based on information entropy optimization not only focus on the representative of the feature item in different categories but also consider the feature distribution in a same category. CHI Statistics based on information entropy optimization is:

$$\chi^2(t_k, c_j)' = \chi^2(t_k, c_j) \times E_{in}(t_k, c_j) \times (1 - E_{\text{out}}(t_k)) \times \lambda_k \tag{8}$$

Here, use the following formula achieve the final value of feature item t_k:

$$\chi^2(t_k) = \max_{1 \le j \le m} \{\chi^2(t_k, c_j)'\} \tag{9}$$

Make a descending order of the evaluation function values of each feature item, select the feature items of top *K* to represent the text information, In order to verify the feasibility and effectiveness of CHI Statistics method based on information entropy optimization, we will adopt the following experiments to validate it.

3. Experiments and Analysis

3.1. Dataset

In order to verify the effectiveness of CHI Statistics text feature selection method, this paper adopts the public data sets. In the course of data preprocessing, all the words are convert to lowercase, remove punctuation, adopt the stop list. Experimental use 10-fold cross-validation method, the data sets used in this experiment as follows:

1)20-Newsgroups data sets, are a data sets widely used in the area of text classification. A total of 19,997 texts in the data sets, and these texts are average assigned to 20 categories.

2) Reuters-21578 data sets, a total of 21,578 articles are Reuters news report. All the news reports are assigned to 135 categories, but each category has one more of news reports. In this experiment we adopt the top 10 categories which contain the most news reports as data sets.

3.2. Evaluation Criteria

In order to evaluate the global classification performance of sorter on all categories, we adopt the usual evaluation methods, Average precision rate of macro ($macro\_p$), Macro average recall ($macro\_r$) and Average F1 value macro ($macro\_F_1$), Calculation formula as follows:

$$macro\_p = \frac{\sum_{i=1}^{|C|} p_i}{|C|} \tag{10}$$

$$macro\_r = \frac{\sum_{i=1}^{|C|} r_i}{|C|} \tag{11}$$

$$macro\_F_1 = \frac{2 \cdot macro\_p \cdot macro\_r}{macro\_p + macro\_r} \tag{12}$$

C represents the number of categories, p_i represents the precision rate of text within category i, r_i represents the Recall rate of category i. p_i represents the probability that the number of text correct classification occupies the number of texts in the sorter. r_i represents the probability that the number of text corrects classification occupy total number of texts.

3.3. Experiment Result

Table 1 list the comparison of the text classification performance of Average F1 value macro, which based on 20-Newsgroups data sets, and run at Bayes classifier [11] and SVM classifier [12], here, Information Gain (IG), CHI Statistics and CHI statistics algorithm based on Information entropy optimization match different quantity of feature items.

Table 1. Average F1 Value Macro Based on 20-Newsgroups Data Sets (%)

Feature number	Bayes classifier			SVM classifier		
	IG	CHI	CHI-IMP	IG	CHI	CHI-IMP
200	44.15	61.63	70.16	53.51	71.92	74.41
400	49.65	70.29	78.05	58.15	76.67	79.91
600	54.38	74.30	80.11	60.30	78.49	81.23
800	57.76	77.26	80.68	62.53	79.07	82.69
1000	59.70	77.74	81.70	64.04	79.66	82.91
1200	62.88	78.46	82.13	65.17	79.26	82.80
1400	64.58	79.01	82.98	66.24	79.11	82.65
1600	65.99	79.58	83.57	67.28	79.18	82.20
1800	67.40	80.04	83.84	67.66	78.60	81.26
2000	68.37	80.61	84.17	68.45	78.62	81.75

For a visual comparison of the Average F1 value macro, we will convert table data to line chart Figure 1 and Figure 2.

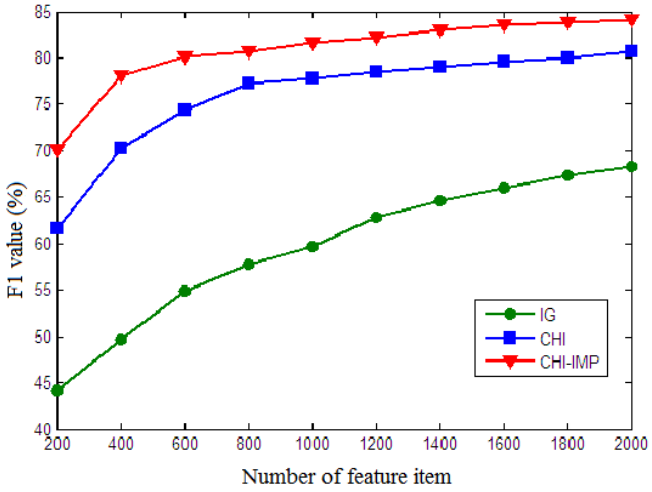

Figure 1. Average F1 Value Macro when Bayesian Classifier Is Running on 20-Newsgroups Data Sets

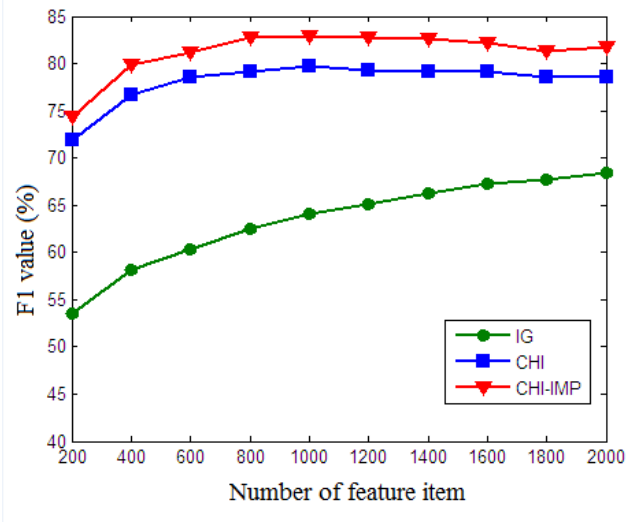

Figure 2. Average F1 Value Macro when SVM Classifier is Running on 20-Newsgroups Data Sets

When using the Bayesian classifier and SVM classifier based on 20-Newsgroups data sets, use Average F1 value macro to measure the effectiveness of text classification, when you select the feature items from 200 to 2000, it can be seen from Figure 1 and Figure 2 that the value of Average F1 of CHI statistics algorithm based on Information entropy optimization is bigger than traditional CHI statistics algorithm, but the IG algorithm always at the lowest status.

In Figure 1, when using Bayesian classifier, with the number of feature items increasing, the Average F1 value macro of traditional CHI Statistical algorithm and CHI Statistics algorithm based on information entropy optimization algorithm also increase, and present a logarithmically increasing trend, that is with the increasing of the feature items number, the increasing trend is decrease. In Figure 2, when using SVM classifier, if there are fewer feature items, with the number of feature items increasing, the Average F1 value macro also increase, when the number of feature items up to 1000, the growth rate will remain unchanged, whereas the Average F1 value macro of IG algorithm always at the lowest status.

We can see from the above analysis, CHI statistics algorithm and CHI statistical algorithm based on information entropy optimization are very similar on classification performance, but, CHI statistics algorithm based on Information entropy optimization algorithm has clearly improved than traditional CHI statistics algorithm under the condition that the Average F1 value macro based on 20-Newsgroups.

In the Reuters-21578 data sets, we also use Bayesian classifier and SVM classifier for processing, when match different number of feature items, IG algorithm, CHI statistics algorithm and CHI statistics based on information entropy optimization algorithm, performance comparison of average F1 value macro, as shown in Table 2 below.

Table 2. Average F1 Value Macro Based on Reuters-21578 Data Sets (%)

Feature number	Bayes classifier			SVM classifier		
	IG	CHI	CHI-IMP	IG	CHI	CHI-IMP
200	59.15	63.76	66.01	62.19	63.08	67.05
400	62.09	64.89	67.97	61.31	62.85	66.59
600	63.86	65.49	68.19	61.73	63.05	67.03
800	64.60	66.02	69.83	62.19	63.12	67.24
1000	64.59	66.69	70.62	62.20	62.72	66.48
1200	64.76	66.96	71.32	62.66	62.96	66.59
1400	64.58	67.06	72.03	62.59	62.88	66.62
1600	65.11	66.92	71.96	62.59	62.93	66.51
1800	65.17	67.15	72.45	62.58	62.84	66.34
2000	65.22	67.35	72.73	62.88	62.73	66.26

For a visual comparison of the three algorithms of average F1 value macro based on Reuters- 21578 data sets, we will convert Table 2 to line chart Figure 3 and Figure 4.

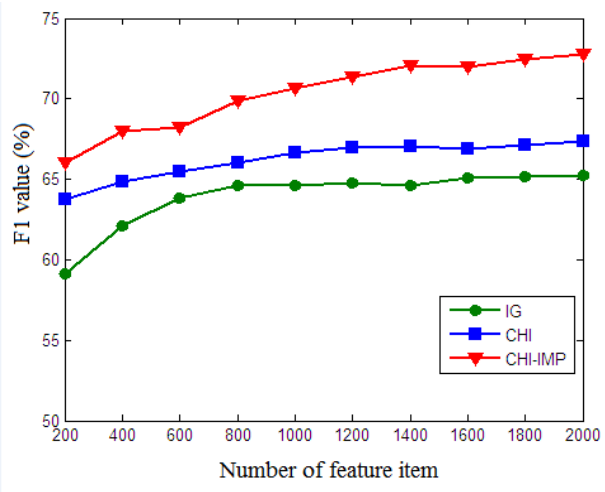

Figure 3. Average F1 Value Macro when Bayesian Classifier is Running on Reuters- 21578 Data Set

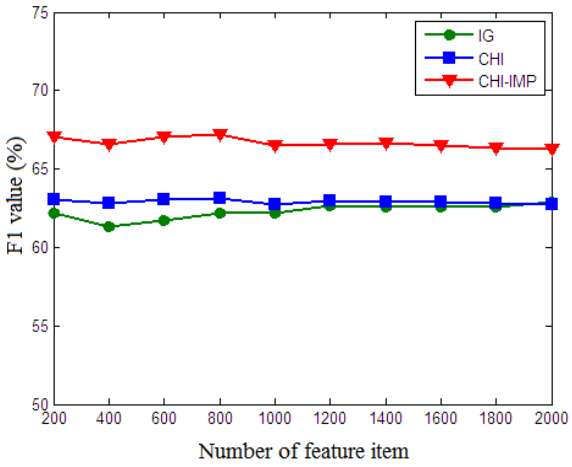

Figure 4. Average F1 Value Macro when SVM Classifier is Running on Reuters- 21578 Data Set

When using the Bayesian classifier and SVM classifier based on 20-Newsgroups data sets, use Average F1 value macro to measure the effectiveness of text classification, when select the number of feature items from 200 to 2000, it can be seen from Figure 3 and Figure 4 that CHI Statistics algorithm based on information entropy optimization algorithm is better than traditional CHI Statistics algorithm on Average F1 value macro.

In Figure 3, when using Bayesian classifier, with the increase of feature items number the Average F1 value macro of IG algorithm, traditional CHI statistical algorithm and CHI statistical algorithm based on information entropy optimization algorithm also increases, the increasing speed is slow. In Figure 4, when using SVM classifier, with the increase of feature items number the Average F1 value macro almost constant. The Average F1 value macro of IG algorithm and CHI Statistical algorithm almost equal.

We can see from the above analysis, in the Reuters-21578 data sets. Average F1 value macro of CHI statistics algorithm based on information entropy optimization is bigger than the traditional CHI statistical algorithm.

4. Conclusion

The traditional CHI statistics algorithm doesn't consider the distribution of feature item frequency within a category or among different categories. This paper proposes a CHI statistics text feature selection method based on information entropy optimization, and introduces information entropy among categories ,information entropy within categories ,feature frequency within a category *etc.* to optimize the existing CHI statistics algorithm. Make the feature frequency is effectively use in the algorithm. Text classification experiment shows that the CHI Statistics text feature selection algorithm based on information entropy optimization has improved the efficiency of text classification.

Acknowledgements

This paper is supported by Key Scientific and Technology Research Project of the State Secrets Bureau (BMKY 2015 S05-1).

References

[1] Y. Xu, W. Xiaodan and Z. Yuxi, "A Survey of Feature Selection", Control and Decision, vol. 27, no. 2, **(2012)**, pp. 161-166.
[2] F. Xueli, F. Haihong and Y. Meng, "Principal component analysis based on mutual information feature selection algorithms", Control and Decision, vol. 28, no. 6, **(2013)**, pp. 915-919.
[3] S. Lili, L. Bingquan and S. Chengjie, "Comparison and Improvement of feature selection in text categorization methods", Journal of Harbin Institute of Technology, vol. 43, no. 1, **(2011)**, pp. 319-324.
[4] Y. M. Yang and X. Liu, "A re-examination of text categorization on methods", Proceeding of the 22nd Annual International Retrieval, New York: ACM, **(1999)**, pp. 42-49.
[5] D. Liuling, H. Heyan and C. Zhaoxiong, "Comparative study on feature extraction method for Chinese text categorization", Journal of Chinese Information Processing, vol. 18, no. 1, **(2004)**, pp. 26-32.
[6] L. Galavotti and F. Sebastiani, "Feature selection and negative evidence in automated text categorization", Proceedings of the ACM KDD-00 Workshop on Text Mining, **(2000)**.
[7] Y. J. Li and C. N. Luo, "Text clustering with feature selection by statistical data", IEEE Transactions. Knowledge and Data Engineering, vol. 20, no. 5, **(2008)**, pp. 641-652.
[8] W. Guang, Q. Yunfei and S. Qingwei, "Feature selection method of collection CHI with the IG", Application Research Of Computers, vol. 29, no. 7, **(2012)**, pp. 2454-2456.
[9] Z. H. Zheng, S. N. Li and G. Zhi, "Multi-label feature selection Algorithm based on information entropy", Journal of Computer Research and Development, vol. 50, no. 6, **(2013)**, pp. 177-184.
[10] P. Yingbo and L. Xiaoxia, "CHI improvement of feature selection in text categorization methods", Computer Engineering and Applications, vol. 47, no. 4, **(2011)**, pp. 128-130.
[11] F. Zheng and I. Geoffrey, "Subsumption resolution: an efficient and effective technique for semi-native Bayesian learning", Machine Learning, New York: ACM, **(1999)**, pp. 42-49.
[12] R. G. Brereton and G. R. Lloyd, "Support vector machines for classification and regression", Analyst, vol. 135, no. 2, **(2010)**, pp. 230-267.

Authors

Guohua Wu, he was born in Jinan, China, on February 24, 1970. Ph.D. professor, major in computer Science and technology, his research interests includes data mining, cryptography, information system, model driven architecture.

Sen Li, he was born in Huaiyang, China, on June 21, 1990. Graduate student, major in computer Science and technology, his research interests includes data mining, cryptography.

Lin Han, she was born in Gongyi, China, on September 15, 1991. Graduate student, major in computer Science and technology, her research interests data mining.

Mengmeng Zhao, she was born in Shenqiu, China, on September 05, 1992. Graduate student, major in computer Science and technology, her research interests includes data mining, cryptography.

