Fuzzy Clustering Algorithm based on Factor Analysis and its Application to Mail Filtering

Jingtao Sun
College of Electrical and Information Engineering, Lanzhou University of Technology, Lanzhou, China
Email: sun2651@126.com

Qiuyu Zhang and Zhanting Yuan
College of Computer and Communication, Lanzhou University of Technology, Lanzhou, China
Email: zhangqy@lut.cn

***Abstract*—Aim at the faults of Dynamic Clustering Algorithm based on Fuzzy Equation Matrix, we raise a fuzzy clustering algorithm based on factor analysis, which it combines the technology of reducing dimension using factor analyses method. The algorithm will deal with the sample collections before fuzzy clustering, which enlarge the scale of using dynamic clustering algorithm to resolve practical problems. All these show that the algorithm has a strong capability of concluding and abstracting through being applied to E-mail filtering. At the same time, we also make an experiment in our optional database. The experiment result verifies that the algorithm recall rate is 87.3 % in the mail filtering, which is higher than the SVM's 80.1%, Naïve Bayes's 61.7%, and KNN's 73.2% respectively. The experiments show that the new algorithm has better recall rate and error rate.**

***Index Terms*—factor analysis, fuzzy clustering, fuzzy equivalence relation, Spam Filtering**

I. PREFACE

Along with the accelerated development of the Internet, the network is, apart from offering us huge amount of information, throwing us to the challenge of better management and classification of such information. As a mathematical method for classification and processing of objects based on their different characteristics, degree of interrelationship, similarity etc, and clustering algorithm is witnessing fast development. Traditional clustering algorithm often makes the assumption that the sample is kind of either or characterized, and the result is the sample is rigidly marked as in a certain category. The fact is that the interrelationship and boundaries of objects are usually not definite in reality, i.e. the interrelationship is fuzzy [1, 2]. In 1965, Zadeh introduced the Fuzzy Set Theory [3], and soon after that in 1969, Ruspinid conducted fuzzy clustering analysis [4-6] based on the concept of fuzzy partition which intends to maximize the similarity among samples classified in the same category and minimize the similarity among those classified in different categories. Among the many fuzzy clustering algorithms, the Dynamic Clustering Algorithm based on Fuzzy Equation Matrix (DCAFEM) [7] proves more practical; however, such algorithms still have some weak points: clustering is seriously time-consuming when processing data in large numbers, and the usual lack of consideration on the interrelationship of characteristic indexes which form each sample in the sample space has led to similar and superfluous characteristic indexes, resulting in increased void sample dimensions, extended clustering time and reduced clustering efficiency.

Aimed at the weak points of common fuzzy clustering algorithms, this paper proposes a fuzzy clustering algorithm based on factor analysis and proceeds to apply it to mail filtering. Tests on customized text corpuses prove that this algorithm achieves higher recall rate and lower false acceptance rate when applied to mail filtering

II. ANALYSIS ON KEY TECHNIQUES

A. Factor analysis

As a technique for dimension reduction and data simplification [8, 9], factor analysis intends to, through analysis on large numbers of original variables, select a few "abstract" variables (common factors) to replace the original ones so that these common factors can be used to represent major information of the original variables, and in this way, the variables can be simplified and variable dimensions can be reduced [10].

The mathematical model can be represented as below [10, 11]:

Suppose there are n original variables, respectively $x_1, x_2, \cdots, x_n$. Since common factors for factor analysis are common influencing factors that cannot be directly observed though in objective existence, so each original variable can be represented by the sum of special factors and the linear function of common factors, i.e. $x_i = a_{i1}F_1 + a_{i2}F_2 + \cdots + a_{it}F_t + \varepsilon_i$, $i = 1, 2, \cdots, n$, wherein $F_1, F_2, \cdots, F_t$ are common factors and ε_i is the special factor of x_i. This model can be represented in matrix: $X = AF + \varepsilon$, where in

© 2009 ACADEMY PUBLISHER

$X=\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}, A=\begin{pmatrix} a_{11} & \cdots & a_{1t} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nt} \end{pmatrix}, F=\begin{pmatrix} F_1 \\ F_2 \\ \vdots \\ F_t \end{pmatrix}, \varepsilon=\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$. Of the model, matrix A is called the factor load matrix, and a_{ij} is called factor "load", the load of the i th variable on the j th factor.

B. Fuzzy clustering analysis

DCAFEM has been introduced in view of the uncertainty of calcification numbers and that dynamic considerations should be made during clustering on the basis of different requirements. It proves relatively practical [4, 12] among all the fuzzy clustering analysis methods, and also is the focal point to be studied in his paper.

Definition 1 (characteristic index matrix) supposes the finite sample set to be classified is $U=\{u_1,u_2,\cdots,u_n\}$ and each sample u_i possesses m characteristic indexes, i.e. u_i can be represented by the m dimension characteristic index vector as $u_i=(u_{i1},u_{i2},\cdots,u_{im})$, $i=1,2,\cdots,n$, wherein u_{ij} stands for the i th characteristic index of the j th sample, and all the characteristic indexes of n samples form the matrix $U^*=\left(u_{ij}\right)_{n\times m}$, and U^* is called the characteristic index matrix of U.

Definition 2 (fuzzy similar matrix) supposes an n fuzzy matrix $R=\left(r_{ij}\right)_{n\times n}$ is on a given finite domain $U=\{u_1,u_2,\cdots,u_n\}$, and $R=\left(r_{ij}\right)_{n\times n}$ is becomes a fuzzy similar matrix if and only if: 1) Reflexivity: $r_{ii}=1$; 2) Symmetry: $r_{ij}=r_{ji}$.

Definition 3 (fuzzy equivalent matrix) supposes an n fuzzy matrix $R'=\left(r_{ij}'\right)_{n\times n}$ is on a given finite domain $U=\{u_1,u_2,\cdots,u_n\}$, and $R'=\left(r_{ij}'\right)_{n\times n}$ is becomes a fuzzy equivalent matrix if and only if: 1) Reflexivity: $r_{ii}'=1$; 2) Symmetry: $r_{ij}'=r_{ji}'$; 3) Transitivity: $R'\circ R'\subseteq R'$.

The operation steps [13, 14] of DCAFEM:

(a) Build the characteristic index matrix for fuzzy clustering analysis;

(b) Data standardization: The dimensions and magnitude orders of m characteristic indexes may vary, so data standardization must be performed for each index value to eliminate the influence caused by variable units and different magnitude orders of the characteristic indexes;

(c) Build the fuzzy similar matrix: When data of $u_{ij}(i=1,2,\cdots,n;\ j=1,2,\cdots m)$ has all been standardized, the degree of similarity between sample $u_i=(u_{i1},u_{i2},\cdots,u_{im})$ and $u_j=(u_{j1},u_{j2},\cdots,u_{jm})$ can be determined through the method of multivariate analysis as: $r_{ij}=R(u_i,u_j)\in[0,1]$, $i,j=1,2,\cdots,n$, and therefore the fuzzy similar matrix $R=\left(r_{ij}\right)_{n\times n}$ s among the samples can be built;

(d) Build the fuzzy equivalent matrix: The fuzzy similar matrix among samples, as built through above steps, may not be necessarily of transitivity, and should be reformed to get a fuzzy equivalent matrix. The specific reform is to build the transitive closure of similarity by obtaining the square of R, that's $R\circ R=R^2$, then $R^2\circ R^2=R^4$, $R^8,R^{16},\cdots$, so on and so forth, until we come to $R^{2K}=R^K$, and R^K is the fuzzy equivalent matrix with which we can proceed with fuzzy clustering analysis.

III. Fuzzy Clustering Algorithm based on Factor Analysis

It can be told through analysis of the abovementioned key techniques that, as a technique for simplification of variables and reduction of variable dimensions, factor analysis makes it possible for us to simplify characteristic indexes in the sample set and reduce the dimensions of sample set. This paper will apply factor analysis to the pretreatment of DCAFEM initial characteristic indexes, in order to expand the capability of DCAFEM in handling practical issues and the efficiency of clustering algorithm in processing problems involving large sample sets.

A. Pretreatment of characteristic indexes

During the analysis of a multi-sample set, the usual situation is that people select a large number of characteristic indexes as an attempt to represent the information of each sample as perfectly as possible and avoid the omission of important information. These many characteristic indexes, however, probably are in close relativity which turns out to complicate sample characteristics, and therefore, factor analysis becomes necessary, in order to eliminate the relativity among samples and reduce the dimensions.

Suppose the finite sample set U comprises n samples, i.e. $u_1,u_2,\cdots,u_n$, each sample u_i composed of m characteristic indexes, that's $U=\left(u_{ij}\right)_{n\times m}=(U_1,U_2,\cdots,U_m)$.

(a) To eliminate the influence caused by variable units and different magnitude orders of characteristic indexes, we need to perform data standardization for each index value.

© 2009 ACADEMY PUBLISHER

Through $\overline{u}_j = \frac{1}{n}\sum_{i=1}^{n} u_{ij}$, ${\sigma_j}^2 = \frac{1}{n}\sum_{i=1}^{n}(u_{ij} - \overline{u}_j)^2$ and $j = 1, 2, \cdots, m$, each row of data can be standardized into $u_{ij}^{'} = \frac{u_{ij} - \overline{u}_j}{\sigma_j}$, $i = 1, 2, \cdots, n$ and $j = 1, 2, \cdots, m$, wherein $\overline{u}_j$ and σ_j respectively stand for the sample average and sample standard deviation of U_j ;

(b) Before factor analysis, similarity among $U_1, U_2, \cdots, U_m$ should be decided through KMO (Kaiser Meyer Olkin, KMO) [15] test to determine the necessity of factor analysis. The KMO statistical magnitude can be any arbitrary value between $(0,1)$; the closer the KMO statistical magnitude is to 0, the weaker the similarity among $U_1, U_2, \cdots, U_m$ is, and closer the KMO statistical magnitude is to 1, the stronger the similarity among $U_1, U_2, \cdots, U_m$ is. We generally believe that factor analysis is can be practically significant when KMO statistical magnitude is bigger than 0.5;

(c) Calculate the covariance matrix of $U_1, U_2, \cdots, U_m$, $\Sigma = (h_{ij})_{m \times m}$; based on $|\Sigma - \lambda I| = 0$, the characteristic equation of Σ , we can get the characteristic root of the covariance matrix as $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$, and the corresponding unit eigenvector is $T_1, T_2, \cdots, T_p$;

(d) According to the principle as applied in handling practical issues, the first t characteristic roots and eigenvectors will be picked out, sum of selected characteristic roots accounting for over 85% of the sum of total characteristic roots, in order to determine the number of common factors;

(e) The factor load matrix $A = (T_1\sqrt{\lambda_1}, T_2\sqrt{\lambda_2}, \cdots, T_m\sqrt{\lambda_m}) = \begin{pmatrix} a_{11} & \cdots & a_{1t} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mt} \end{pmatrix}$ can be calculated through the characteristic roots and eigenvectors of Σ . If the load of each factor on different characteristic indexes does not vary distinctively, the factor load matrix should be rotated, usually through the orthogonal rotation method, to get $A^{'} = \begin{pmatrix} b_{11} & \cdots & b_{1t} \\ \vdots & \ddots & \vdots \\ b_{m1} & \cdots & b_{mt} \end{pmatrix}$;

(f) Through computing of row vectors of the rotated factor load matrix $A^{'}$, under conditions that $b_{ip} = Max\{b_{i1}, b_{i2}, \cdots, b_{it}\}$, $i = 1, 2, \cdots, m$ and $p \in \{1, 2, \cdots, t\}$. By retaining the maximum load value b_{ip} of U_i in t factors in matrix $A^{'}$, we get matrix $A^{*} = (b_{ij}^{'})_{m \times t}$, $i = 1, 2, \cdots, m$; $j = 1, 2, \cdots, t$, where in $b_{ij}^{'} = \begin{cases} b_{ip}, & j = p \\ 0, & other \end{cases}$;

(g) Through the operation steps mentioned above, the finite sample set U can be simplified into the finite sample U^{Δ} which consists of n samples, each sample u_i composed of t characteristic index factors; therefore we get the characteristic index matrix $U^{*} = (u_{ij}^{*})_{n \times t}$ of n samples, wherein u_{ij}^{*} stands for the j th characteristic index factor of the i th sample, and the computing formula be:

$$u_{ij}^{*} = \sum_{q=1}^{m} u_{iq} b_{qj} , \ i = 1, 2, \cdots, n; \ j = 1, 2, \cdots, t \qquad (1)$$

B. Build the fuzzy similar matrix

After pretreatment of characteristic indexes, the sample set can be clustered, and we need to apply the method of multivariate analysis to determine the similarity between sample u_i and sample u_j : $r_{ij} = R(u_i, u_j) \in [0,1], i, j = 1, 2, \cdots, n$. In this way, we can build a fuzzy similar matrix $R = \begin{pmatrix} r_{11} & \cdots & r_{1n} \\ \vdots & \ddots & \vdots \\ r_{n1} & \cdots & r_{nn} \end{pmatrix}$; the computing formula shall be:

$$r_{ij} = \frac{1}{t}\sum_{k=1}^{t} \exp\{-\frac{3}{4}(\frac{u_{ik}^{*} - u_{jk}^{*}}{\sigma_k})^2\} \qquad (2)$$

Wherein ${\sigma_k}^2 = \frac{1}{n}\sum_{i=1}^{n}(u_{ik}^{*} - \overline{u}_k)^2$, $\overline{u}_k = \frac{1}{n}\sum_{i=1}^{n} u_{ik}^{*}$.

The fuzzy similar matrix $R = \left(r_{ij}\right)_{n \times n}$ built through the abovementioned method may not necessarily be of transitivity, so we must transform the fuzzy similar matrix R into fuzzy equivalent matrix R^{*} before the clustering of sample set U , and proceed to perform dynamic clustering based on the fuzzy equivalent matrix. This paper shall apply the fuzzy transitive closure method to make clustering, in order to increase the universality of algorithms, as specified below:

© 2009 ACADEMY PUBLISHER

(a) Use the square method to seek the transitive closure $t(R)$ of the fuzzy similar matrix R, i.e. $R^2 \Rightarrow R^4 \Rightarrow \cdots \Rightarrow R^{2k} = t(R)$, wherein $k \le [\log_2 n]+1$;

(b) Select a proper confidence level value $\lambda \in [0,1]$ to seek the λ cut matrix $t(R)_\lambda$ of $t(R)$. When λ varies between $[0.1]$, the corresponding classification changes accordingly.

C. Algorithm description

a) Input the initial data of finite sample set U;
b) Output sample clustering results.

S1) Standardize initial data to generate the characteristic index set $U_1, U_2, \cdots, U_m$;

S2) Shrink $U_1, U_2, \cdots, U_m$ through factor analysis and build the characteristic index matrix U^* of the sample set;

S3) Calculate the fuzzy relation among sample data to generate fuzzy similar matrix R;

S4) Seek the transitive closure $t(R)$ of R;

S5) Select a proper parameter λ to seek the λ cut matrix $t(R)_\lambda$ of $t(R)$;

S6) Output the data of each cluster.

IV. Application of the Fuzzy Clustering Algorithm Based on Factor Analysis in Mail Filtering

Major research approaches to mail filtering are currently based on statistics, i.e. the acquisition of the different characteristics between spam mails and valid mails through statistical studies on existing mails, in order to realize effective distinction of incoming mails based on matching of these characteristics. As we make intensive studies on methods such as artificial intelligence, machine learning etc, the focus of researches has been shifted to the application of these methods in mail filtering, for example, the mail filtering algorithm based on support vector machine, the mail filtering algorithm based on neural network, the mail filtering algorithm based on rough set etc. Most of these filtering methods, however, have the problem of low recall rate and high false acceptance rate.

In view of the current lack of strong distinction characteristics and similar characterization etc between many spam mails and valid mails, this paper, by referring to fuzzy clustering, a significant method for solving uncertainty problems, proposes the fuzzy clustering algorithm based on factor analysis for mail filtering purposes, and tests on customized text corpuses prove that this algorithm achieves higher recall rate and lower false acceptance rate when applied to mail filtering.

A. Analysis on algorithm application

One of the problems pressing for solution before we can apply the fuzzy clustering algorithm based on factor analysis to mail filtering is the selection of a proper parameter λ, for different λ values in fuzzy clustering algorithm directly lead to different clustering results, and this exactly is the enchantment of dynamic clustering algorithm. Through adjustment of parameter λ, we can make the result of mail filtering as satisfactory as possible. However, when the value of λ is too small, categories after clustering of mail sample sets become too few with over high degree of abstraction, leading to reduced recall rate and increased false acceptance rate, and on the contrary, if the value of λ is too big, categories after clustering of mail sample sets become too many, resulting in seriously degradation of filtering efficiency, though we get higher recall rate and lower false acceptance rate. This paper tends to apply the F -statistic method to obtain a proper λ value based on which we can seek categories that lead to comparatively optimized recall rate and false acceptance rate when applied to mail filtering.

The fuzzy clustering algorithm based on factor analysis divides the spam sample set into several subcategories, and the center point of each category can be obtained through definition 1 and the arithmetic process as stated below:

Within the set U^Δ, suppose Q is the category quantity of the corresponding λ value, n_Q as the sample quantity of the i th category, samples of i th category being $A_1^i, \cdots, A_{n_Q}^i$, $A_J^i \in \{u_1, u_2, \cdots, u_n\}$, $J = 1, 2, \cdots, n_Q$, the clustering central vector of the i th category is $O^i = (\overline{O}_1^i, \overline{O}_2^i, \cdots, \overline{O}_t^i)$, wherein $\overline{O}_k^i$ stands for the average of this category of samples at the k th characteristic index factor, $\overline{O}_k^i = \frac{1}{n}\sum_{j=1}^{n_i} u_{ik}$, $k = 1, 2, \cdots, t$.

By calculating the distance between the characteristic index vectors (indexes generated after pretreatment of characteristic indexes) of new mails and the center point of the category, mail filtering can realize identification of spam mails on the basis of set comparison threshold, and in this way the search band can be narrowed to improve the efficiency of mail filtering.

B. Simple numerical examples

We explain the effect of adopting fuzzy clustering algorithm based on factor analysis in the mail filtering by a simple test set. In terms of the less document amount, the weight information of the words cannot be expressed. We will not consider the weight problem of the words in document here. When comparing the similarity of different words, correlation factor is used as measurement. New here we have 6 documents, 3 of them are ads on Internet money-making and the rest 3 are porno ads, as indicated in Table 1:

© 2009 ACADEMY PUBLISHER

TABLE. 1.
ORIGINAL DOCUMENT

Cont	Doc
Mail1	与其天天耗在网上闲聊玩游戏看美女写真,不如与我一道在网上边娱乐边赚钱,比闲聊泡美女强多了!
Mail2	让你在上网娱乐的同时也有一份不错的收入.二十一世纪,网上赚钱不在是神话,边游戏边赚钱成为可能.娱乐赚钱两不误
Mail3	网上挂 QQ 泡美女不赚钱.挂这个可以赚钱哦!只要能上网成为会员就能有收入.只要你努力.月收入几千不是梦!
s-mail1	爱我美女为您提供日本美女,性感美女,美女写真, 淫荡视频,绝对都是让你喷火的美女.点击进入观看。
s-mail2	太漂亮了,上网观看淫荡美女写真《漂亮日本美女写真集》《性感韩国美女写真》观看的美女视频和发表观看感言。
s-mail3	寂寞的美女在黄色网站淫荡的呼风唤雨，要想观看露的最多的激情视频美女,美女写真,欢迎点击进入成为会员。

Extract 12 key works from the 6 documents to build the term-document matrix X , as shown in Table 2.

TABLE. 2.
WORD-DOCUMENT ORIGINAL MATRIX

Word	Mail1	Mail2	Mail3	s-mail1	s-mail2	s-mail3
网上	2	1	1	0	0	0
闲聊	2	0	0	0	0	0
游戏	1	1	0	0	0	0
娱乐	1	2	0	0	0	0
赚钱	1	3	2	0	0	0
上网	0	1	1	0	0	0
收入	0	1	2	0	0	0
美女	2	0	1	5	4	3
观看	0	0	0	1	3	1
写真	1	0	0	1	3	1
视频	0	0	0	1	1	1
淫荡	0	0	0	1	1	1

Table 3 is the correlation matrix of the 12 characteristic indexes, and it's perceptible that the matrix contains many relatively high correlation coefficients.

TABLE.3.
CORRELATION MATRIX OF THE 12 CHARACTERISTIC INDEXES

	网上	闲聊	游戏	娱乐	赚钱	上网	收入	美女	观看	写真	视频	淫荡
网上	1	.8783	.9258	.7143	.378	0	-.1429	-.4472	-.5112	-.5112	-.6547	-.6547
闲聊	.8783	1	.6325	.2928	0	-.3162	-.2928	-.1309	-.3492	-.3492	-.4472	-.4472
游戏	.9258	.6325	1	.9258	.6124	.25	0	-.6211	-.5522	-.5522	-.7071	-.7071
娱乐	.7143	.2928	.9258	1	.7559	.4629	.1429	-.7028	-.5112	-.5112	-.6547	-.6547
赚钱	.378	0	.6124	.7559	1	.9186	.7559	-.9297	-.6763	-.6763	-.866	-.866
上网	0	-.3162	.25	.4629	.9186	1	.9258	-.8281	-.5522	-.5522	-.7071	-.7071
收入	-.1429	-.2928	0	.1429	.7559	.9258	1	-.7028	-.5112	-.5112	-.6547	-.6547
美女	-.4472	-.1309	-.6211	-.7028	-.9297	-.8281	-.7028	1	.6858	.6858	.8783	.8783
观看	-.5112	-.3492	-.5522	-.5112	-.6763	-.5522	-.5112	.6858	1	1	.7809	.7809
写真	-.5112	-.3492	-.5522	-.5112	-.6763	-.5522	-.5112	.6858	1	1	.7809	.7809
视频	-.6547	-.4472	-.7071	-.6547	-.866	-.7071	-.6547	.8783	.7809	.7809	1	1
淫荡	-.6547	-.4472	-.7071	-.6547	-.866	-.7071	-.6547	.8783	.7809	.7809	1	1

Table 4 is the proposed factor variance contribution table, where the 2 extracted factors are listed from top to bottom by the magnitude of variance contributions, and it's perceptible that the fire 2 factors are already capable of interpreting 86.982% of the variance of the original characteristic indexes and embody most of the information.

TABLE.4.
TOTAL VARIANCE EXPLAINED

Cont	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	7.668	63.903	63.903	7.668	63.903	63.903	6.120	50.996	50.996
2	2.770	23.079	86.982	2.770	23.079	86.982	4.318	35.986	86.982
3	.953	7.941	94.924						
4	.513	4.272	99.196						
5	.097	.804	100						
6	2.92E-016	2.43E-015	100						
7	6.62E-017	5.52E-016	100						
8	-1.97E-018	-1.64E-017	100						
9	-3.80E-017	-3.16E-016	100						
10	-1.31E-016	-1.10E-015	100						
11	-2.17E-016	-1.81E-015	100						
12	-3.04E-016	-2.53E-015	100						

Table 5 shows the characteristic index factor matrix generated after characteristic index treatment and it can be seen that the original term-document matrix has been effectively simplified with reduced dimensions.

© 2009 ACADEMY PUBLISHER

TABLE.5.
FACTOR MATRIX

	Factor	
	1	2
Mail1	2.464	5.577
Mail2	0	3.555
Mail3	0.884	0.815
s-mail1	7.414	0
s-mail2	9.314	0
s-mail3	5.646	0

Table 6 gives the clustering result obtained through DCAFEM, and it's enough to show that clustering of mails by the fuzzy clustering algorithm based on factor analysis is correct, and the subsequent comparison operation is feasible.

TABLE 6
CLUSTERING RESULT OF DCAFEM

1	Mail1 Mail2 Mail3
2	s-mail1 s-mail2 s-mail3

C. Example analysis

In analysis of the performance of mail filtering, selection of language material library is extremely important. There are now some authoritative standard language material libraries overseas, such as PU1 [16] language material library. Yet in China, an authoritative standard language material library is nowhere to be found. In this case, we've prepared 2,561 spam mails extensively collected to build a training set A which contains 2,341 spam mails and a probe set B which contains 220 spam mails. Since the formation mode of samples in the mail sample set is similar to the formation mode of characteristic indexes of samples in simple numerical examples, and to intensify the significance of characteristic indexes of samples under different circumstances, we generally need to introduce a weight value for representation purposes. This paper has introduced the characteristic term frequency- sample frequency in samples as the weight, and data standardization can be undone since mail sample sets do not have such questions as varied units and magnitude orders.

To test the effectiveness of the algorithm put forward in this paper, we use Matlab7.4 to write emulator programs on an IBM ThinkPad T43p, with an Intel Pentium M Dothan Processor 2.13GHz CPU and a 2G SDRAM memory to respectively realize the mail filtering algorithm of SVM [17, 18], the mail filtering algorithm of Naïve Bayes [19] and the mail filtering algorithm of KNN [20], all algorithms experimented on customized sets. Training shall be done first on the training set A to get various clustering results and build the spam mode library based on which we can proceed with identification of spam mails on the probe data set B.

Evaluation of spam identification is mainly based on the following 3 indexes:

(a) Recall rate, i.e. the detection rate of spam, reflects the ability of spotting spam mails; higher recall rate means less "unfiltered" spam.
(b) Precision rate, i.e. the rate of right judgment of spam, reflects the ability of "finding" spam mails; bigger precision rate means fewer valid mails will be misjudged as spam mails.
(c) F_1 value, i.e., the harmonic mean of the recall rate and the precision rate, actually integrates the recall rate and the precision rate as one judgment index.

The formulas are as following:

Recall: $R = \frac{N_A}{N_S} \times 100\%$;

Precision: $P = \frac{N_A}{N_A + N_B} \times 100\%$;

F value: $F_1 = \frac{2RP}{R+P} \times 100\%$.

Where N_A is number of spam which are judged correctly; N_S is number of actual spam; N_B is number of normal email that are judged as spam.

Result of the above experiment is shown in Table 7. The experiment has realized the anticipated effects and has verified the feasibility and superiority of this algorithm, offering a new approach to the research of mail filtering algorithm.

TABLE 7
COMPARISON OF EXPERIMENTAL RESULTS OF THE FUZZY CLUSTERING ALGORITHM BASED ON FACTOR ANALYSIS, SVM, NAÏVE BAYES AND KNN

	Recall/ (%)	Precision/ (%)	F_1 value / (%)	Execute time/ s
Naïve Bayes	61.7	89.53	73.05	0.76
KNN	73.2	80.21	76.54	1.03
SVM	80.1	86.4	83.13	1.53
fuzzy clustering algorithm based on factor analysis	87.3	92.37	89.76	1.34

VI. CONCLUSION

The fuzzy clustering algorithm based on factor analysis applies factor analysis to simplify the characteristic indexes on sample sets, and manages to retain the information of the original sample sets while greatly simplifying the information, making the subsequent fuzzy clustering analysis more maneuverable; through DCAFEM, samples after pretreatment of characteristic indexes can be clustered and the center point of each category can be calculated. Then comparison between the new sample and the center point of the category helps identification of mails, which can increase the precision of mail filtering and the ability of identifying unknown spam mails. Like other clustering algorithms, this algorithm needs to be improved in several aspects, such as its astringency and effectiveness, selection of parameter λ, definition of similarity relation etc.

REFERENCES

© 2009 ACADEMY PUBLISHER

[1] Lei Y.J, Sun J.P, Wang B.S. On the Fussy Knowledge Processing and Extensions of Fuzzy Sets theory [J]. Journal of Air Force Engineering University (Natural Science Edition). 2004, 5(3): 40-44.
[2] Lu X, Liao J.M. Study of Software Quality Evaluation Based on Fuzzy Sets Theory [J]. Journal of University of Electronic Science and Technology of China. 2007, 36(3): 652-655.
[3] Kang S.W, Wang Y.M. An Approach to Generating Rules Based on Rough and Fuzzy Sets Theories [J]. Journal of Xiamen University (Natural Science). 2002, 41(2): 173-176.
[4] Xie W.X, Gao X.B. The development of Fuzzy clustering theory and its application [J]. Chinese Journal of Stereology and Image Analysis. 1999, 4(2): 113-119.
[5] Liu H, Huang S. A Genetic Semi-supervised Fuzzy Clustering Approach to Text Classification [J]. Artificial Intelligence and Soft Computing - ICAISC 2004. 2003: 173-180.
[6] Bordogna G, Pagani M, Pasi G. A Dynamic Hierarchical Fuzzy Clustering Algorithm for Information Filtering [J]. Neural Information Processing. 2006: 3-23.
[7] Chen D.F, Lei Y.J, Tian Y. Clustering Algorithm Based on Intuitionistic Fuzzy Equivalent Relations [J]. Journal of Air Force Engineering University(Natural Science Edition). 2007, 8(1): 63-65.
[8] Li X.R. Compare and Application of Principal Component Analysis, Factor Analysis and Clustering Analysis [J]. Journal of Shandong Education Institute. 2007, 22(6): 23-26.
[9] Holgado-tello F, Carrasco-ortiz M, Del B.G. Factor analysis of the Big Five Questionnaire using polychoric correlations in children[J]. Quality and Quantity. 2007, 9(01):85-89.
[10] Shi H.B, Lu Y.L. Research on the Effect of Factor-Analysis-Based Dimension Reduction on Classification Performance [J]. Journal of North University of China(Natural Science Edition). 2007, 28(6): 556-561.
[11] Niu X.Q, Chen L. Research on Chinese Partition of Manufacture Industry Based on Factor Analysis [J]. Journal of Wuhan University of Technology (Social Sciences Edition). 2007, 20(6): 792-795.
[12] Dong Y.Y. Fuzzy cluster analysis [J]. Journal of Jing gang shan University. 2006, 27(06): 26-28.
[13] Xie J.G. Swatch Classification based on fuzzy Equivalence Relation [J]. Journal of Yun cheng University. 2006, 24(2): 10-11.
[14] Qu X, Zhang Z.F. An Advanced Fuzzy Cluster Analysis Algorithm Based on Equivalence Relation [J]. Journal of Lanzhou Jiaotong University. 2003, 22(3): 17-20.
[15] Guo Y, Gu H.Y. The Factorial Analysis of the Determinants of Capital Structure [J]. Journal of Shandong Normal University (Natural Science). 2008, 23(2): 4-6.
[16] Androu T.I, Pal Iouras G.M.K.E. Learning to Filter Unsolicited Commercial E-mail[R]. http://www.aueb.gr/users/ion/docs/TR2004_updated.pdf.
[17] Ai-bing J, Jia-hong P, Shu-huan L, et al. Support Vector Machine for Classification Based on Fuzzy Training Data[C]. 2006 International Conference on Machine Learning and Cybernetics, Dalian, pp: 1609-1614, 2006.
[18] Basu A, Walters C, Shepherd M. Support vector machines for text categorization[C]. System Sciences, Proceedings of the 36th Annual Hawaii International Conference on, pp: 7-11, 2003.
[19] Androu T.I, Pal Iouras G.M.K.E. An Evaluation of Naïve Bayesian Anti-Spam Filtering[C]. The 11 European Conferences on Machine Learning (ECML 2000).pp:221-225
[20] Lishuang L, Lishuang L, Tingting M, et al. Extracting location names from Chinese texts based on SVM and KNN [C]. Natural Language Processing and Knowledge Engineering, Proceedings of 2005 IEEE International Conference on. pp: 371- 375.

Jingtao Sun Doctor student. Born in Daqing Heilongjiang province in 1981, have published many academic papers in domestic core magazine and international conference. His research interests include: information security, Chinese text classification, Anti-Spam etc.

Qiuyu Zhang Associate professor and master tutor. Vice dean of School of computer and communication in Lanzhou University of Technology, director of software engineering center, vice dean of Gansu manufacturing information engineering research center, director of "software engineering" characteristic research direction and academic group of Lanzhou University of Technology. His research interests include: image processing and pattern recognition, multimedia information processing, information security, software engineering etc.

Zhanting Yuan Professor and doctor tutor in the School of Computer and Communication Engineering, Mr Yuan received his MASc in June 1989 from Artificial Intelligent and Robot Research Center of Xi'an Jiaotong University. Now Professor Yuan is the Science Leader of computer science and technology, director of Gansu manufacturing information engineering research center, administrative director of Chinese electrical higher education, and the first batch of Chinese new century "Bai qian wan" person with ability. The research interests of professor Yuan include: image processing and pattern recognition, computer vision, Software engineering, information security etc.

© 2009 ACADEMY PUBLISHER