Journal of Theoretical and Applied Information Technology
20th February 2013. Vol. 48 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

THE RESEARCH ON EMAIL CLASSIFICATION BASED ON Q-GAUSSIAN KERNEL SVM

LIFAN[1], MA TAO[2], XU HONG[1]

[1]School of Computer Science and Technology, Chengdu University of Information Technology, Chengdu 610225, China

[2]Teachers' College of Beijing Union University, Beijing 100011, China

ABSTRACT

The use of different kernel functions in SVM (Support Vector machines) has been reported in the literature. In this paper, the use of the q-Gaussian function as kernel function in SVM is investigated and q-Gaussian function is explored. While the q-Gaussian kernel SVM classifiers being built, cross validation methods are used to select the non-extensive entropic index q under varying feature sizes, the punishment parameter and kernel width are set empirically, the email classification on two leading Chinese email corpuses, TREC06c and CCERT, are implemented with SVM classifiers employing Gaussian kernel and q-Gaussian kernel. Experiment results show that q-Gaussian kernel SVMs can enhance the classification performance effectively.

Keywords: *E-mail Classification, Q-Gussian Kernel, SVM, Classification Performance*

1. INTRODUCTION

Email is an efficient, rapid and cheap mean of communication in modern society. With the popularization, low cost and fast delivery, there have been a great amount of unsolicited bulk email messages, referred to spam emails. Spam emails have been a severe problem on the Internet, increasing in exponential speed all the time. Spam could give rise to the consumption of computer and network resources, the cost of human time and attention in dismissing unwanted messages and information security, so they are frustrating and annoying. Combating spam is a difficult job, automatic e-mail filtering seems to be the most effective method. Usually, spam filtering take the approaches of content analysis, which can be regarded as a special case of text categorization with target values being spam and legitimate (non-spam).

Statistical modeling is a more efficient and promising content-based spam filtering technology. Generally, there are two different approaches to spam classification: generative models (for instance, Naive Bayesian classifier [1]) and discriminative models. Discriminative models includes Support Vector Machines (SVM [2, 3]), and Logistic Regression (LR) [4]. SVM motivated by statistic learning theory[5], is a powerful classification technique in data mining and has been successfully applied to many real-world applications due to excellent generalization performance and a unique ability to handle extremely large feature spaces (such as text). SVM performs classification by finding the optimal hyperplane that maximizes the margin between the two classes. With nonlinearly separable data, SVM employs kernel trick to cope. The data points are mapped from input space to high dimensional feature space where linear discrimination is possible, the inner product in feature space can be carried out with kernel function in input space. Kernel functions must satisfy certain criteria known as Mercer conditions, and be selected according to different applications. Many researchers devote themselves to kernel construction and application. There are many kernel functions available, including linear, polynomial, Gaussian kernel and other kernel functions, among which Gaussian kernel is more suitable [6] and reasonable choice for SVM classification [7].

In non-extensive statistical mechanics there exists q-Gaussian function, which parameterizes standard Gaussian function by replacing exponential expression with q-exponential expression[8], while maximizing Tsallis entropy [9] under certain constraints [10,11] . Moreover, the q-Gaussian function can continuously and smoothly reproduce different radial basis functions, like the Gaussian,

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

the Inverse Multiquadratic, and the Cauchy functions, by changing a real parameter q. The q-Gaussian function has been introduced into RBF neural networks as the radial basis function [12, 13]. In this paper the q-Gaussian function is proposed to be incorporated into SVM as kernel function, resulting in q-Gaussian kernel SVM, which is employed for email classification on real email datasets, the classification performance compared with SVM based on Gaussian kernel. The proof about availability of q-Gaussian function is also presented.

This paper contains five sections. Following Section 1, the theory and technology about SVM are reviewed in Section 2. Section 3 explores the q-Gaussian function and discusses the availability of q-Gaussian function as kernel function. Section 4 presents experiments on email classification with q-Gaussian kernel SVM, the results be compared and analyzed. Finally, the conclusion of this study is drawn in Section 5.

2. SUPPORT VECTOR MACHINE (SVM)

2.1 Support Vector Classification

In content-based email classification, emails are represented with Vector Space Model (VSM), which is widely used in text mining. An email, which generally containing the body, the subject and other header fields, can be processed and represented as a high dimensional sparse vector. Each vector component value denotes the contribution of the corresponding word (or term) from the bag-of-words with respect to the certain email. With huge amount of email features, among which many are irrelevant and redundant, classification model should be trained with more calculation complexity, and has poor generalization ability. Hence high dimensional data should be processed with some dimension reduction methods including feature selection and feature extraction. After that email classifier should be constructed, here Support Vector Machine is taken as the classification algorithm.

SVM has been reported as a classification model with remarkable performance on text categorization task [2, 3, 14], which has many special advantages in solving finite data size; nonlinear and high-dimensional machine learning problems. With high-dimensional feature spaces, most traditional techniques such as nearest-neighbors and neural network, fail due to the "curse of the dimensionality" because they are based on the minimization of the empirical risk. Differently, SVM operates on another induction principle, called structural risk minimization principle from statistical learning theory. SVM can overcome the problem of over fitting and local minimum and gain better generalization capability. Kernel trick is incorporated into SVM, which doesn't increase the computational complexity, furthermore overcomes the curse of dimensionality problem effectively. The incorporation of positive definite kernel into SVM can be interpreted as an embedding of the input space into a high dimensional feature space where the classification is carried out without using explicitly this feature space. Hence, the problem of choosing architecture for a neural network application is replaced by the problem of choosing a suitable kernel for a Support Vector Machine. In the following, the simple description about the theory and implementation of SVM classification algorithm is provided.

Considering linear separable classification with a training set of N two-class data points $D=\{\mathbf{x}_i, y_i\}_{i=1}^{N}$, the input vector $\mathbf{x}_i=\left(x_i^{(1)}, x_i^{(2)}, \cdots x_i^{(n)}\right)\in \Re^n$, and the target label $y_i\in\{-1,+1\}$. SVM tries to find an optimal separating hyperplane $(\mathbf{w}\ \mathbf{x})+b=0$, which separates all the training samples correctly as much as possible. That means that the following condition should be satisfied:

$$y_i\left[(\mathbf{w}\ \mathbf{x}_i)+b\right]\geq 1, i=1,2,\cdots N \tag{1}$$

Equation (1) comes down to the construction of two parallel bounding hyperplanes at opposite sides of the separating hyperplane $(\mathbf{w}\ \mathbf{x})+b=0$ with the margin width between two bounding hyperplanes equals to $\frac{2}{\|\mathbf{w}\|^2}$. The optimal hyperplane SVM finds is the one whose margin is the largest, which can be found by solving the following constrained optimization problem:

$$\underset{\mathbf{w},b}{Min}\ \frac{1}{2}\mathbf{w}^T\mathbf{w}\ \text{ s.t. } y_i\left[(\mathbf{w}\ \mathbf{x}_i)+b\right]\geq 1, \nabla i \tag{2}$$

It is called the primal problem. To solve it, introduce lagrange multiplier $\alpha_i\geq 0$, there exists a Lagrange equation:

$$L(\mathbf{w},b,\boldsymbol{\alpha})=\frac{1}{2}\mathbf{w}^T\mathbf{w}-\sum_{i=1}^{N}\alpha_i(y_i\left[(\mathbf{w}\ \mathbf{x}_i)+b\right]-1) \tag{3}$$

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

With Wolfe theory the problem can be transformed to its dual problem:

$$\max .W(\boldsymbol{\alpha}) = \sum_i \alpha_i - \frac{1}{2}\sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle \quad s.t. \ \ \alpha_i \geq 0, \sum_i \alpha_i y_i = 0 \tag{4}$$

This is a quadratic programming (QP) problem, which can be solved with optimization methods. If $\boldsymbol{\alpha}$ found, $\mathbf{w}$ and b can be calculated with $\mathbf{w} = \sum_{i=1}^{l} \alpha_i y_i \mathbf{x}_i$. For the novel data $\mathbf{x}$, the class is determined with:

$$y(\mathbf{x}) = \text{sgn}((\mathbf{w}\ \mathbf{x}) + b) \tag{5}$$

Training vectors that satisfy $y_i\left[(\mathbf{w}\ \mathbf{x}_i) + b\right] = 1$ (the corresponding Lagrange multipliers non-zero) are support vectors, which determine the optimal hyperplane. All other training examples are irrelevant for defining the binary class boundaries.

In the case of linearly non-separable training data, slack variables are introduced, the primal problem is as following:

$$\underset{\mathbf{w},b}{Min}\ \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_i \zeta_i \quad \text{s.t. } y_i\left[(\mathbf{w}\ \mathbf{x}_i) + b\right] \geq 1 - \zeta, \zeta \geq 0, \nabla i \tag{6}$$

It's seen that SVM finds the best compromise between the complexities of the model and learning ability, implied by the above two items respectively. With nonlinearly separable data, data are mapped form input space to feature space, where the maximal margin classifier can be built. Avoiding the map function φ represented explicitly, the kernel function $K(\mathbf{x}_i, \mathbf{x}_j) = \langle \varphi(\mathbf{x}_i), \varphi(\mathbf{x}_j) \rangle$ in input space can replace the inner product in feature space. For the kernel function, the common choices contain the linear kernel $K(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^T \mathbf{x}_j$, the polynomial kernel $K(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i^T \mathbf{x}_j + 1)^d$ and Gaussian kernel $K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2})$ ($\sigma \in \Re^+$, constant). The resulting SVM classification model can be written as:

$$y(\mathbf{x}) = \text{sgn}(\sum_{i=1}^{N} \alpha_i y_i K(\mathbf{x}, \mathbf{x}_i) + b) \tag{7}$$

It is known that kernel functions must satisfy certain criteria for preserving the convexity of the problem.

2.2 Mercer Kernel

Mercer Theorem tells, if the kernel K is a symmetric positive definite function, which satisfies the following conditions:

$$K(\mathbf{x}_k, \mathbf{x}) = \sum_i^{\infty} a_i \phi_i(\mathbf{x}_k)\phi_i(\mathbf{x}), \ a_i > 0 \tag{8}$$

And

$$\iint K(\mathbf{x}_k, \mathbf{x}) g(\mathbf{x}_k) g(\mathbf{x}) d\mathbf{x}_k d\mathbf{x} > 0 \tag{9}$$

Then the kernel K would represent an inner product in feature space $K(\mathbf{x}_k, \mathbf{x}) = \phi(\mathbf{x}_k)\ \phi(\mathbf{x})$ (10)

and is known as Mercer Kernel. Mercer condition needs to be satisfied for keeping the problem convex and hence obtaining a unique solution. In next section q-Gaussian function will be explored and can be used as kernel function in SVM.

3. THE Q-GAUSSIAN FUNCTION

The Gaussian distribution is ubiquitous in probability and statistics, serving as an attractor of independent systems with finite variance. It maximizes the Boltzmann-Gibbs entropy under appropriate constraints:

$$S_{BG} = -\int_{-\infty}^{\infty} \ln[p(x)]p(x)dx \tag{11}$$

The q-Gaussian is not an alternative to the classic Gaussian function but a parametric generalization of Gaussian function, because it reproduces Gaussian function when $q \to 1$. As the probability distribution function, the q-Gaussian function arises naturally when the central limit theorem from sum of random variables with global correlations is considered [15, 16]. It's closely related with the non-extensive q-statistics and the generalized q-entropy. The q-generalization of the classic entropy (equation 11) as the basis for generalizing the classic theory reaches its maximum at the distributions usually referred to as q-Gaussian. The q-Gaussian distribution arises as an attractor of certain correlated systems, which maximizing the non-extensive entropy:

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

$$S_q = \frac{1-\int_{-\infty}^{\infty}[p(x)]^q dx}{q-1} \quad ,q \in \Re \tag{12}$$

under appropriate constraints, $\mu_q = \frac{\int x[p(x)]^q dx}{\int [p(x)]^q dx}$, $\sigma_q^2 = \frac{\int (x-\mu_q)^2 [p(x)]^q dx}{\int [p(x)]^q dx}$ and $\int p(x)dx = 1$, where μ_q is the q-mean, defined analogously to the usual mean, except using the so-called q-expectation values, σ_q^2 is defined analogously to the usual second order central moment.

With q-Gaussian function, q-deformed logarithm and its inverse (q-exponential) should be known,

$$\ln_q(x) = \frac{x^{1-q}-1}{1-q} \quad , x > 0 \tag{13}$$

Its first derivation is $\frac{d}{dx}\ln_q(x) = \frac{1}{x^q}$. It's seen that the deformed logarithm is always a strictly increasing function because the derivative is positive for any value of q. The q-exponential is the inverse function of $\ln_q(x)$:

$$e_q(x) = [1+(1-q)x]_+^{1/(1-q)} \tag{14}$$

Where $[x]_+ = \max\{0, x\}$, the q-exponential functions are always positive when $1 < q < 3$. These functions reduce to the usual logarithm and exponential functions when $q \to 1$. The q-Gaussian density is defined for $-\infty < q < 3$ as

$$p(x;\mu_q,\sigma_q) = A_q\sqrt{B_q}\left[1-(1-q)B_q(x-\mu_q)^2\right]_+^{1/(1-q)} = A_q\sqrt{B_q}e_q^{-B_q(x-\mu_q)^2} \tag{15}$$

A_q can be derived using density function normalization $\int p(x)dx = 1$ [15]. The width of the distribution is characterized by $B_q = [(3-q)\sigma_q^2]^{-1}$ $,q \in (-\infty, 3)$. An interesting property of the q-Gaussian function is that it can reproduce different RBFs (Radial Basis Function) for different values of the real parameter q. The q-Gaussian distribution has a compact form for $q < 1$ and decays asymptotically as a power law for $1 < q < 3$. The second order moment is finite for $q < 5/3$. For $q > 3$ the distribution cannot be normalized. The q-Gaussian distribution reduces to the usual Gaussian distribution $p(x) = \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ when $q \to 1$. The q-Gaussian distribution equals to the Cauchy distribution $p(x) = \frac{\sigma}{\pi((x-\mu)^2+\sigma^2)}$ when $q = 2$. The standard q-Gaussian $Z \quad N_q(0,1)$ has zero mean and unit q-variance, its density is written as $p(x;0,1) = \frac{A_q}{\sqrt{3-q}}\left[1-\frac{q-1}{3-q}x^2\right]_+^{\frac{1}{1-q}}$. Figure 1 presents some standard probability density functions, including Gaussian, Cauchy, and q-Gaussian with other values of non-extensive entropic index q.

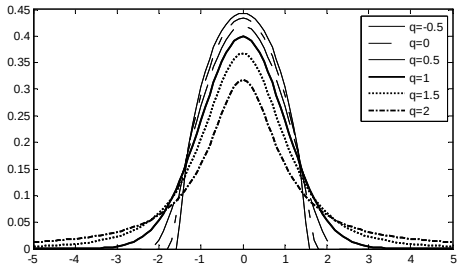

Fig. 1. q-Gaussian Probability Density Function ($q = -0.5, 0, 0.5, 1, 1.5, 2$)

The q-Gaussian functions will be incorporated into SVM, using for email classification in this paper. The q-Gaussian functions satisfies Mercer conditions, the q-Gaussian kernel function be written as

$$K(\mathbf{x}_i, \mathbf{x}_j) = e_q\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{(3-q)\sigma_q^2}\right) \tag{16}$$

4. EXPERIMENTS AND RESULT

The Experiments are carried out with Chinese email corpuses, and the SVM classifier be implemented based on LIBSVM library [20]. Email classification performance with q-Gaussian kernel

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

SVM is evaluated, and compared with other kernel functions.

4.1 Performance Metrics

In order to evaluate email classification performance with q-Gaussian kernel SVM, some metrics should be used. Usually the accuracy, precision and recall rates are used. These measures are calculated using confusion matrix given below:

Table 1. Confusion Matrix For Email Classification

	Spam	Legitimate Email
Classified as Spam	a	b
Classified as Legitimate	c	d
Assuming that there are N emails to be tested, $N = a+b+c+d$		

Accuracy (A) of email classification is calculated by dividing the number of correctly classified samples by the total number of test samples and is defined as: $A = \frac{a+d}{N}$, it represents classification rate with all tested emails.

Precision (P) measures the system's ability to present only relevant items while recall (R) measures system's ability to present all relevant items. $P = \frac{a}{a+b}$, the higher the precision rate, the less the misclassification of legitimate emails is. $R = \frac{a}{a+c}$, the higher the recall rate, the less the misclassification of spam is. With email classification, the precision is more important than the recall rate, because a legitimate email shouldn't be discarded, otherwise the user may suffer from loss.

4.2 Data Sets

Two widely used Chinese email corpuses, TREC06c [17] and CDSCE (CCERT Data Sets of Chinese Emails) [18] are used in our experiments. TREC06c corpus is a public Chinese email corpus coming from the Text Retrieval Conference (TREC), it consists of 64620 messages, 21766 of which are marked as spam and 42854 are tagged as ham. CCERT Data Sets are made up of 63710 emails, including 18314 legitimate emails and 45396 spam messages. The spam messages were collected via using honey pot technique (SPAMPOT) and the ham email are from the Chinese public forums. The information about the corpus is shown in Table 2.

Table 2. Statistic Of The Experimental Corpus

Corpus	language	Spam	Ham	total
TREC06C	Chinese	42854	21766	64620
CCERT	Chinese	45396	18314	63710

With the two corpuses, the contents of email subject and body (the two important constitutions of email) are only extracted, the messages are preprocessed with a Chinese term segmentation tool ICTCLAS [19] developed by Chinese Academy of Sciences. Finally those emails are represented with the bag of word technique.

4.3 Results

Two experiments are performed on two public corpuses TREC06C and CDSCE respectively. SVM classifier is built incorporating Gaussian kernel and q-Gaussian kernel with some values of non-extensive entropic index q. With two email corpus data, the χ^2 statistic (CHI) method is used for feature selection to reduce data dimensions to obtain varying dimension data. Two parameters in SVM with Gaussian kernel should be tuned, including punishment parameter C and kernel parameter $\delta = 1/2\sigma^2$, decided through trial and error in our experiments. It is found effective that C is set 30, δ is set 0.01 with the two email data. In q-Gaussian kernel SVM, the punishment parameter has the same value, δ_q (that is $1/\sigma_q^2$) is also set 0.01. With respect to non-extensive entropic index selection, cross validation is employed to determine.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

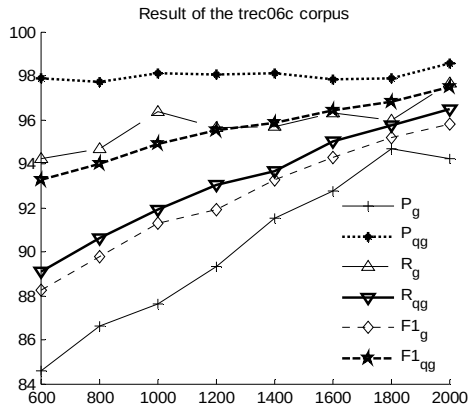

Fig 2. Average Metric Values On TREC06c Corpus With 10-Fold Cross Validation

In experiments with TREC06C, 6000 emails are randomly selected for the dataset, which have spam-to-ham ratio of 2:1, the same ratio as original dataset. The feature sizes are set from 600 to 2000 with a step of 200. For each dimension, the 5-fold cross validation is implemented to select the parameter q with the best performance. Based on the previous process, the ten-fold cross validation (10-CV) tests are used in all the data for each dimension. With the best q values available in each dimension data, the average classification results of precision, recall and F1 on the trec06c corpus are shown in Figure 2.

Table 3 presents Gaussian kernel SVM email classification performance metrics including accuracy means, three standard deviations about accuracy, precision and recall rate, from 10-fold cross validation tests with varying feature sizes. In the same way, Table 4 provides q-Gaussian kernel SVM email classification performance metrics including accuracy means, three standard deviations about accuracy, precision and recall rate, from 10-fold cross validation tests with varying feature sizes and the corresponding best q values. It's seen that he two models both have improvements on precision with the feature size increasing. The q-Gaussian SVM reaches its best precision 98.56% when the feature size is 1400, and for Gaussian kernel achieves its highest precision 94.68% if the feature size is 1800 (As illustrated in Fig. 2). Furthermore, the q-Gaussian kernel SVM models obviously achieve better performance than the RBF kernel SVM models. Generally, the two models have some rise in recall and F1 metrics with the feature size increasing. In F1 metric, the q-Gaussian kernel models have distinguished improvements over the RBF kernel model, thought they fail to defeat the RBF kernel model in recall rate.

Table 3. Results On TREC06C (10-Ford Validation, Gaussian Kernel, Varying Feature Sizes)

	600	800	1000	1200	1400	1600	1800	2000
ACC	82.78	85.30	87.58	88.58	90.58	92.04	93.38	94.20
std_ACC	7.14	5.75	4.84	3.99	2.40	1.66	1.81	1.54
std_P	12.64	11.05	9.14	8.90	7.02	5.65	4.54	4.13
std_R	4.95	4.35	4.61	3.87	3.96	3.46	2.58	2.56

Table 4. Results On TREC06C (10-Ford Validation, Q-Gaussian Kernel, Varying Feature Sizes And The Best Q Values)

	600	800	1000	1200	1400	1600	1800	2000
q	1.5	1.5	1.5	1.5	1.6	1.6	1.5	2.9
ACC	91.26	92.18	93.26	93.98	94.44	95.18	95.74	96.62
std_ACC	0.87	0.81	0.66	0.75	0.51	0.47	0.71	0.77
std_P	1.01	0.93	0.80	0.61	0.83	1.06	1.36	1.05
std_R	1.79	1.58	0.99	1.06	1.04	0.74	0.74	1.08

ACC is accuracy mean in the ten-ford validation experiment, std_ACC is accuracy standard deviation. std_P and std_R are Precision and Recall standard deviations .

From Table 3 and 4, it is know that the q-Gaussian models have better ACC with each feature size than the Gaussian kernel model. What's more, the former one also have a lower standard deviation on the accuracy, precision and recall, which indicates that the q-Gaussian kernel SVMs have steady and superior performance compared with the Gaussian kernel SVMs.

In experiment with CDSCE corpus, 4000 emails are chosen at random with spam to legitimate ratio of 1:1. Here the processed data have different dimensions (they are 800, 1200, 1600, 2000, 2200, 2400, 2600 and 2800), the experimental processes are similar to above mentioned. The results are given in Figure 3, Table 5 and 6. As shown in the figure and table, the two classification models all

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

have high precision rates over 96%, the top average precision with q-Gaussian classifier is 99.70% while the counterpart metric is 98.69% with the Gaussian kernel classifier. It's the same that the Q-Gaussian kernel models are also superior to the RBF kernel models. Further more, the Q-Gaussian kernel models also shows outstanding performance on the F1, and for the recall rates, the two models are more or less the same.

According to the two above tables, the q-Gaussian kernel classifiers have a little improvement over Gaussian kernel classifier except with the feature size of 2000. The q-Gaussian kernel classifiers have a low std_ACC and std_R when the feature size is less than 2000, but the two metrics increase a little. In std_P, the q-Gaussian models are better than the Gaussian models.

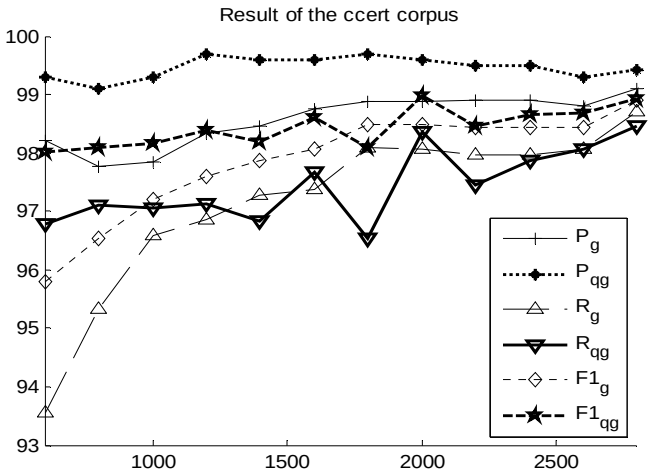

Fig 3. Average Metric Values On CCERT With 10-Fold Cross Validation

Table 5. Results On CDSEC (10-Ford Validation, Gaussian Kernel, Varying Feature Sizes)

	800	1200	1600	2000	2200	2400	2600	2800
ACC	96.65	97.65	98.1	98.5	98.45	98.45	98.45	98.9
std_ACC	2.06	1.34	1.02	0.77	0.65	0.65	0.57	0.62
std_P	1.97	1.28	1.01	0.93	1.03	1.03	0.97	0.81
std_R	3.04	1.85	1.61	1.12	1.08	1.08	0.91	0.80

Table 6. Results On CDSEC (10-Ford Validation, Q-Gaussian Kernel, Varying Feature Sizes And The Best Q Values)

	800	1200	1600	2000	2200	2400	2600	2800
q	2.8	2.8	2.2	1.8				
ACC	98.1	98.45	98.25	98.15	98.5	98.7	98.7	98.95
std_ACC	0.89	0.76	0.63	0.39	0.67	0.71	0.64	0.65
std_P	1.21	0.46	0.49	0.49	0.67	0.67	0.89	0.78
std_R	0.94	1.62	1.29	1.01	1.47	1.45	1.12	1.19

5. CONCLUSIONS

In this paper, a new spam filtering approach is presented, using Support Vector Machine with q-Gaussian function as kernel function. Originating in non-extensive statistical mechanics, q-Gaussian function has better flexibility, being used in many research fields. Email classification experiments on two leading Chinese corpuses, TREC06c and CCERT are taken with q-Gaussian kernel SVMs and Gaussian kernel SVMs. The results demonstrate that the new model outperforms the SVM with Gaussian kernel in precision, recall and accuracy rate. The conclusion can be drawn that SVM with q-Gaussian kernel might be a useful filtering algorithm that can separate spam from legitimate

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

email efficiently, and can be introduced to solve other machine learning problems.

ACKNOWLEDGEMENTS

This work was supported by National Statistics Scientific Research Program of China (2010LC14) and the Scientific Research Foundation of CUIT (CSRF201003).

REFERENCES

[1] Youn, S. a. A Comparative Study for Email Classification. JOURNAL OF SOFTWARE, 2 (3), 1-13, 2006.

[2] H. Drucker, D. Wu and V. N. Vapnik. Support Vector Machines for Spam Categorization. IEEE Transactions on Neural Networks, 10(5), 1048-1054, September, 1999.

[3] Amayri, O. and Bouguila, N. A study of spam filtering using support vector machines. In Proceedings of Artif. Intell. Rev. 73-108, 2010.

[4] H. Yong, Y. Muyun, Q. Haoliang, H. Xiaoning, and L. Sheng. The Improved Logistic Regression Models for Spam Filtering. Asian Language Processing, 2009. IALP '09. International Conference on, 2009, pp. 314-317.

[5] V. Vapnik, Statistical Learning Theory. New York: John Wiley & Sons, 1998.

[6] Yu-Yen Ou, Chien-Yu Chen, Shien-Ching Hwang, Yen-Jen Oyang, Expediting Model Selection for Support Vector Machines Based on Data Reduction, In IEEE Proc. SMC, pp. 786-791, 2003.

[7] Hsu, C.W., Chang, C.C., Lin, C.J.. A Practical Guide to Support Vector Classification. Department of Computer Science, National Taiwan University, http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf

[8] R. Tinos, L. Murta, Selection of radial basis functions via genetic algorithms in pattern recognition problems, in: Tenth Brazilian Symposium on Neural Networks, 2008. SBRN '08, pp. 171–176, 2008.

[9] C. Tsallis, Possible generalization of Boltzmann–Gibbs statistics, Journal of Statistical Physics 52 (1–2) pp. 479–487, 1988.

[10] C. Tsallis, R.S. Mendes, A.R. Plastino, The role of constraints within general- ized nonextensive statistics, Physica A: Statistical Mechanics and its Applica- tions 261 (3–4), pp.534–554, 1998.

[11] R. Dı´az, E. Pariguan, On the Gaussian q-distribution, Journal of Mathematical Analysis and Applications 358 (1) (2009) 1–9.

[12] Wei Zhao, Ye San: RBF neural network based on q-Gaussian function in function approximation. Frontiers of Computer Science in China 5(4), pp.1-6, 2011.

[13] Tinós, R. and Júnior, L.O.M. Use of the q-Gaussian Function in Radial Basis Function Networks. In Proceedings of Foundations of Computational Intelligence (5), pp. 127-145, 2009.

[14] Shu-wei chih, Tsan Ying Yu. Email Spam Filtering using SVM with selected Kernel Parameters, 2009 fourth international conference on innovative computing, information and control.

[15] Thistleton W, Marsh J A, Nelson K, Tasllis C. Generalized Box-Müller method for generating q-Gaussian random deviates. IEEE Transactions on Information Theory, 53(12): 4805–4810, 2007.

[16] Umarov, S., Tsallis, C., Steinberg, S.: On a q-central limit theorem consistent with nonextensive statistical mechanics. Milan Journal of Mathematic (2008), doi: 10.1007/s00032-008-0087-y.

[17] TREC 2006 Spam Track Public Corpora: http://plg1.cs.uwaterloo.ca/cgi-bin/cgiwrap/gvcormac/foo06

[18] CERNET Computer Emergency Response Team (CCERT) http://www.ccert.edu.cn/spam/sa/ datasets. htm

[19] The Chinese segmentation tool ICTCLAS can be download form: http://ictclas.org/

[20] Chih-Chung Chang and Chih-Jen Lin, LIBSVM : a library for support vector machines. http://www.csie.ntu.edu.tw/~cjlin/libsvm