IETE Journal of Research

ISSN: 0377-2063 (Print) 0974-780X (Online) Journal homepage: http://www.tandfonline.com/loi/tijr20

Chi-square Statistics Feature Selection Based on Term Frequency and Distribution for Text Categorization

Chuanxin Jin, Tinghuai Ma, Rongtao Hou, Meili Tang, Yuan Tian, Abdullah Al-Dhelaan & Mznah Al-Rodhaan

To cite this article: Chuanxin Jin, Tinghuai Ma, Rongtao Hou, Meili Tang, Yuan Tian, Abdullah Al-Dhelaan & Mznah Al-Rodhaan (2015) Chi-square Statistics Feature Selection Based on Term Frequency and Distribution for Text Categorization, IETE Journal of Research, 61:4, 351-362, DOI: 10.1080/03772063.2015.1021385

To link to this article: https://doi.org/10.1080/03772063.2015.1021385

Chi-square Statistics Feature Selection Based on Term Frequency and Distribution for Text Categorization

Chuanxin Jin[1], Tinghuai Ma[1,2], Rongtao Hou[1], Meili Tang[3], Yuan Tian[4], Abdullah Al-Dhelaan[4] and Mznah Al-Rodhaan[4]

[1]School of Computer and Software, Nanjing University of Information Science & Technology University, Nanjing, China, [2]Jiangsu Engineering Centre of Network Monitoring, Nanjing University of Information Science & Technology, Nanjing, China, [3]School of Public Administration, Nanjing University of Information Science & Technology, Nanjing, China, [4]Computer Science Department, King Saud University, Riyadh, Saudi Arabia

ABSTRACT

Text categorization (TC) becomes the key technology to find relevant and timely information from a volume of digital documents, and feature selection techniques are proposed to overcome the high dimensionality which causes the high computational complexity and low accuracy in TC tasks. Chi-square statistics (CHI) is one of the most efficient feature selection methods; however, it has two weaknesses. (1) It is document frequency based, and only counts whether the term occurs or not. Actually, high-frequency term occurring in few documents is often regarded as a discriminator in corpus. (2) It does not consider the term distribution. A term has more discriminating power for a specific category when its difference in degree of distribution is lower. In this paper, we propose a modified CHI feature selection approach which is called term frequency and distribution based CHI to overcome these weaknesses. We use sample variance to calculate the term distribution, and improve the classic CHI with maximum term frequency. Extensive and comparative experiments on three corpora show that the proposed approach is comparable to the classic feature selection methods in terms of macro-F1 and micro-F1.

Keywords:
Chi-square statistics, Difference degree of distribution, Feature selection, Term frequency, Text categorization.

1. INTRODUCTION

Text categorization (TC) is defined as assigning new unlabelled documents to a set of pre-defined categories based on classification patterns [1,2,3]. The volumes of digital documents available online are growing exponentially. TC becomes a key technology to find relevant and timely information from these documents for many applications [4], such as customer relationship management [5], spam email filtering [6,7], web page classification [8], text sentinel classification [9], software bug classification [10], etc.

TC is naturally treated as a supervised learning problem, and several algorithms from machine learning (ML) approaches have been used as TC classifiers in the past years, such as *k*-nearest neighbour (*k*NN) [11], support vector machines [12], and Naive Bayes [13,14]. In recent years, many other classifiers are proposed by researchers. Zhang et al. [15] propose a novel projected-prototype based classifier, in which a document category is represented by a set of prototypes, each assembling a representative for the documents in a subclass and its corresponding term subspace. Nguyen et al. [16] propose an improved centroid-based classifier which uses the precise term-class distribution properties instead of the presence or absence of terms in classes.

A major problem which makes TC different from other classification tasks is the high dimensionality of the feature space due to a large number of terms. This problem increases the computational complexity of ML methods used for TC and brings about inefficiency as well as low accuracy due to redundant or irrelevant terms in the feature space [3,11,17,18]. Therefore, feature selection (FS) techniques are proposed to reduce the high dimensionality under the premise of guaranteeing the performance of classifiers.

From document frequency perspective, the classical methods almost use document frequency (DF) sufficiently, but we also cannot ignore the impact of term frequency (TF) [19]. For instance, two words having TFs of 10 and 100, respectively, and DFs are both 10, which means that we are unable to judge their relative importance according to DF; on the other hand, TF considers such information which may be useful in the selection of important features.

In this paper, we propose a modified chi-square statistics (CHI) FS approach called term frequency and distribution based CHI (TFDCHI). Then, we do experimental verification with *k*NN classifiers on common corpora. The rest of this paper is organized as follows: in Section 2, we describe the related work about FS, and analyse

the drawbacks of four classic FS methods; in Section 3, we propose our modified chi-square statistics FS method based on term frequency and the term distribution; in Section 4, we describe the data-set, classifier, and performance measures used in our experiments; Section 5 presents the experimental results and shows the effectiveness of our new approach; and finally, we conclude this paper with future work.

2. RELATED WORK

FS is a process that selects a subset from the original feature set according to some criteria of feature importance [3,20]. There are two major ways of viewing FS [21]. The first one is wrapper approach that selects the term subset using the evaluation function which act as a wrapper around the classifier algorithm, and these features will be used on the same classifier algorithm [22]; the other is the filter approach that selects the feature subset from the original feature space using one evaluation function which is independent to the classifier algorithm [22]. Since the filter FS approach is simple and efficient, it has been widely used in the TC, such DF [11,23,24], CHI [11,23,24], information gain (IG) [11,23,24], mutual information (MI) [11,25], and expected cross-entropy (ECE) [26]. The IG and CHI are the two most efficient feature selections, and DF is comparable with the performance of IG and CHI [11]. In this paper, our new approach is also based on filter selection. We will give detailed definitions on these classical methods. Formally, for a specific category, we give the following definition used in [24,27]:

a is the number of documents with term and belongs to category

b is the number of documents with term and do not belongs to category

c is the number of documents without term and belongs to category

d is the number of documents without term and do not belongs to category

Thus, N is the number of documents in training set, $N = a + b + c + d$.

2.1 The Basic Principle of Chi-square Statistic and Improvement

The CHI is used to measure the lack of independence between t_i and C_k, and the basic principle of chi-square is the thing which determines whether the hypothesis is true by observing the deviation of the actual value and the theoretical value. The specific steps are given as follows:

1. Assume that two variables are independent.
2. Observe the deviation of the actual value and the theoretical value.
3. According to the deviation, decide to accept the original assumption or the alternative hypothesis.

In step 3, if the deviation is small enough, we can argue that this error is caused by imprecise measuring methods or accident. Actually, two variables are independent; therefore, we accept the original assumption. However, if the deviation is big enough, we consider that it is unlikely to be accidental or caused by inaccurate measurement, and then we accept the alternative hypothesis.

The CHI formula is used to calculate the deviation of the actual value and the theoretical value. Assume that a term t_i and a category C_k are independent, and the related variables are shown in Table 1.

Because of the independence of t_i and C_k, the term t_i should occur equiprobably in all documents, so this probability is defined as follows:

$$p(t_i) = \frac{a+b}{N} \tag{1}$$

According to Table 1, whether a document containing a term t_i belongs to category C_k has four circumstances. The number of documents that belongs to category C_k are $a + c$; therefore, in these documents, the number of documents where term t_i should occur is given by

$$N_1 = (a+c) \times p(t_i) \tag{2}$$

Then, according to Eq. (2), we can calculate the deviation of the actual value and the theoretical value for the documents with term t_i which belongs to category C_k, the formula is shown as follows:

$$D_1 = \frac{(a - N_1)^2}{N_1} \tag{3}$$

Similarly, the deviation of the remaining circumstances can be calculated and indicated as $D_2, D_3, and\, D_4$. Finally, the deviation of term t_i and category C_k is

Table 1: Related variables

	Belong to category C_k	Do not belong to category C_k	Sum
The number of documents with term t_i	a	b	$a + b$
The number of documents without term t_i	c	d	$c + d$
Sum	$a + c$	$b + d$	$N = a + b + c + d$

calculated by

$$\chi^2(t_i, C_k) = D_1 + D_2 + D_3 + D_4 \tag{4}$$

This is the CHI formula, and then we simplify Eq. (4) and obtain the formula as follows:

$$\chi^2(t_i, C_k) = \frac{N \times (ad - bc)^2}{(a+c) \times (b+d) \times (a+b) \times (c+d)} \tag{5}$$

If $\chi^2(t_i, C_k) = 0$, the term t_i and the category C_k are independent; therefore, the term t_i does not contain any category information. Otherwise, the greater the value of the $\chi^2(t_i, C_k)$, the more category information the feature t_i owns.

The score of term t_i in text collection is obtained by averaging or maximizing the category-specific scores:

$$\chi^2_{\text{avg}}(t_i) = \sum_{k=1}^{|C|} p(C_k)(t_i, C_k) \tag{6}$$

$$\chi^2_{\max}(t_i) = \max_{k=1}^{|C|}\{(t_i, C_k)\} \tag{7}$$

In real-world corpus, considering the presence or absence of a term at the same time, the performance of CHI is better than that of MI. The weakness of CHI method is that it exaggerates the role of the low-frequency term and ignores the term distribution.

Many improved methods for CHI have been proposed in recent years, such as Dai et al. [28] focused on the redundant features and proposed an algorithm which based on modified CHI feature selection and rough set. This approach can evidently improve classification precision. Yunus and Khan [29] proposed the bivariate non-central chi-square distribution by compounding the Poisson probabilities with bivariate central chi-square distribution. They discussed both equal and unequal degrees of freedom. Xia et al. [30] and Evgeniy et al. [31] found that the term distribution was an important discriminator and should be considered in TC. Cerchiello and Giudici [32] used a non-parametric approach with the combination of classification tree which is constructed by using chi-square test to improve online text classification.

Meanwhile, there were many other publications using CHI, such as Chen and Chen [33] proposed using CHI to measure similarities and chi-square tests to determine the homogeneity of two random samples of term vectors for TC. They found that the combination of CHI and cosine similarities can provide a more reliable and effective categorization mechanism. Lee and Luh [34] presented an inverse chi-square based web content classification system. They classified bilingual web pages at an average precision rate of 97.11%, and maintained a favourably low false-positive rate. Tian et al. [35] proposed a fuzzy TOPSIS (Technique for Order Preference by Similarity to an Ideal Solution) model for information source selection. In this model, the Euclidean distance is substituted by the value of chi-square test to refine the relative closeness.

2.2 Classic Methods

There are many classic methods in feature selection for TC.

2.2.1 Document Frequency (DF)

DF is a simple FS method, and it computes the number of documents where a term t_i occurs. The basic idea is that the rare terms are not useful for category prediction and may degrade the global performance [11]. In spite of its simplicity, it has a performance similar to IG and CHI if the keyword number is not too low. It is defined as follows:

$$\text{DF}(t_i, C_k) = p(t_i \mid C_k) \tag{8}$$

where $(t_i \mid C_k) = a$.

2.2.2 Mutual Information (MI)

MI is an important concept in the information theory and it measures the mutual dependence of the term and the category. The weakness of MI is that the score is strongly influenced by the marginal probabilities of terms, because rare terms will have a higher score than common terms [27]. The approximate formula is

$$\text{MI}(t_i, C_k) = \log \frac{a \times N}{(a+c) \times (a+b)} \tag{9}$$

Lee and Kim [36] propose an FS method that derives from MI between selected features and the label set; this method improves the classification performance to a great extent.

2.2.3 Information Gain (IG)

IG measures the information content that the presence or absence of a term to determine whether the document belongs to a category; it is from entropy in information theory. The weakness of the IG method is that it prefers to select terms distributed in many categories, but these terms have less discriminating power in TC

tasks. It is calculated by

$$\begin{aligned} \mathrm{IG}(t_i) = & -\sum_{k=1}^{|C|} p(C_k)\log p(C_k) \\ & + p(t_i)\sum_{k=1}^{|C|} p(C_k \mid t_i)\log p(C_k \mid t_i) \\ & + p(\overline{t}_i)\sum_{k=1}^{|C|} p(C_k \mid \overline{t}_i)\log p(C_k \mid \overline{t}_i) \end{aligned} \tag{10}$$

where $p(C_k) = (a+c)/N$, $p(t_i) = (a+b)/N$, $p(\overline{t}_i) = (c+d)/N$, $p(C_k \mid t_i) = a/(a+b)$, $p(C_k \mid \overline{t}_i) = c/(c+d)$.

Shang et al. [37] proposed a novel metric called global information gain which can avoid redundancy naturally and an efficient FS method called maximizing global information gain; the new method has better results than others methods in most cases and run faster than the traditional higher order algorithms. Uguz [38] proposed a two-stage FS method by combining IG, principal component analysis, and genetic algorithm; the experimental results show that the proposed model is able to achieve high categorization effectiveness.

2.2.4 Expected Cross-Entropy (ECE)

Finally, ECE is similar to IG, but it only considers the terms occurred in a document and ignores the absent terms [26]. It can be defined as

$$\mathrm{ECE}(t_i) = p(t_i)\sum_{k=1}^{|C|} p(C_k \mid t_i)\log\frac{p(C_k \mid t_i)}{p(C_k)} \tag{11}$$

where $p(C_k) = (a+c)/N$, $p(C_k \mid t_i) = a/(a+b)$.

3. MODIFIED CHI-SQUARE STATISTICS APPROACH BASED ON TERM FREQUENCY AND DISTRIBUTION

3.1 A Toy Example

In order to illustrate the weakness of CHI method, a hypothetical training set (Table 2) is composed of three categories and two terms, and each category has 10 documents. The DF and TF of a term are presented, respectively, as follows.

In Eq. (5), in the case of the values of b, c, and d are invariability, the value of $\chi^2(t_i, C_k)$ will be greater with the growth of a. Meanwhile, the value of a is only associated with the DF which includes t_i in a category C_k and the TF of t_i in a category C_k is ignored. Now in Table 2, the value of DF of t_1,t_2 is 10 and 9, respectively; therefore, according to Eq. (5), the relevance of t_1 and C_1 is more closely and the term t_1 will be selected as the feature. However, actually, we can observe that the value of TF of t_2 is 72, which are much higher than the value of t_1 (10); this indicates that the term t_2 has well discriminability and should be chosen.

Table 2: A hypothetical training set

	Category C_1	Category C_2	Category C_3
Document frequency of t_1	10	0	2
Document frequency of t_2	9	2	0
Term frequency of t_1	10	0	2
Term frequency of t_2	72	2	0

This toy example clearly indicates one weakness of CHI method which exaggerates the role of the low-frequency term. To be specific, CHI will be likely to choose more terms whose DF is high and TF is low as feature. Some terms whose TF is high but DF is low will be ignored and it may bring about inefficiency and low accuracy. Actually, these terms are what we really need for TC.

According to Eq. (5), the evaluation function is composed of five variables (*a, b, c, d, N*). Therefore, another weakness of CHI method is that it did not consider the term distribution. Some research works [30,31,39] have shown that terms which have relative uniform distribution in a single category should have more discriminability than those of uneven distribution in the same category.

3.2 TFDCHI Feature Selection Method

To deal with the problems described in Section 3.1, we introduce a modified CHI FS approach called TFDCHI. This method adds TF and term distribution to the traditional CHI method. In this section, we introduce the whole process of the new approach and explain why this new approach can choose feature effectively.

After the process of pre-processing, documents are typically represented by vector space model (VSM). The content of a text is represented by a vector in the term space. Therefore, we can obtain a training set $D = \{d_1, d_2, \ldots, d_N\}$, $d_j = \{w_1, w_2, \ldots, w_{|T|}\}$, and a category set $C = \{c_1, c_2, \ldots, c_{|C|}\}$, where $d_j (1 \le j \le N)$ is a document of set D, $w_i (1 \le i \le |T|)$ is the weight of a term $t_i (1 \le i \le |T|)$ in one document, $c_k (1 \le k \le |C|)$ is a category of the corpus, N is the number of documents in training set, $|T|$ is the number of terms in training set, and $|C|$ is the number of categories in the training set.

In this paper, to make up for the weakness of traditional CHI method, first, we introduce the TF.

Now, tf_{ij} is the term frequency of a term t_i in the document d_j. If we consider the term frequency tf_{ij} is a sample, and in a category c_k, we can obtain a sample of

value set $\mathrm{TF}=\{tf_{i1}, tf_{i2}, \ldots, tf_{iN_k}\}$.Then, we can obtain the average term frequency $\overline{tf}_i$ and the maximum term frequency $tf_{i\max}$ as follows:

$$\overline{tf}_i = \frac{1}{N_k}\sum_{j=1}^{N_k} tf_{ij}, \tag{12}$$

$$tf_{i\,\max} = \max_{j=1}^{N_k}\{tf_{ij}\}, \tag{13}$$

where N_k is the number of documents in the category c_k. If a term occurs a lot of times in a document of a specific category, but this kind of document is very rare, unfortunately, this term will not be chosen as a feature since the traditional CHI method is based on DF. Because of this, we will put TF into the CHI formula.

Second, we introduce the term distribution. The definition of the term distribution as follows.

In a specific category, if a feature is distributed evenly, then the feature has higher associated value with the category. That is to say, if a feature has relative uniform distribution, then the term distribution of this feature is lower, and the amount of classified information of this feature is greater; therefore, this feature has more discriminating power for this specific category.

In this paper, we use sample variance to calculate the difference of term distributions. In the traditional science, sample variance is one of the statistics which is used commonly and a metric which is used to measure the variation or dispersion degree of a data-set. According to Eq. (8), the average term frequency $\overline{tf}_i$ is the sample mean, and the sample variance is calculated as follows:

$$V(t_i, c_k) = \frac{1}{N_k - 1}\sum_{j=1}^{N_k}(tf_{ij} - \overline{tf}_i)^2 + \alpha \tag{14}$$

$\alpha = 0.0001$ is a very small real number. Here it is the term distribution.

Finally, we give two approaches based on the TF and term distribution to calculate the value in category c_k as follows:

$$\mathrm{THDCHI1}(t_i, c_k) = \frac{\log(1 + tf_{i_{\max}}) \times \chi^2(t_i, c_k)}{V(t_i, c_k)}, \tag{15}$$

$$\mathrm{TFDCHI2}(t_i, c_k) = \frac{tf_{i_{\max}} \times \chi^2(t_i, c_k)}{V(t_i, c_k)}. \tag{16}$$

In the TFDCHI1(t_i, c_k), we use logarithmic function to smooth the effect of the maximum term frequency $tf_{i\,\max}$, then based on the basic theory of chi-square, the greater the value of CHI, the more category information the feature owns. In the second approach, we delete the logarithmic function to make the value of the formula as larger as possible. In the above two formulas, we use the maximum TF rather than the average TF to calculate the value, because we need to reserve the terms which have strong discriminability as the TF is high but the DF is relative low.

In the corpus, the score is obtained by averaging or maximizing the category-specific scores. Here we also use the maximum way to obtain the value, because the higher value of a term in a special category, the more discriminability the term owns for this category. The formula is as follows:

$$\mathrm{TFDCHI1}_{\max}(t_i) = \max_{k=1}^{|C|}\{\mathrm{TFDCHI1}(t_i, c_k)\}, \tag{17}$$

$$\mathrm{TFDCHI2}_{\max}(t_i) = \max_{k=1}^{|C|}\{\mathrm{TFDCHI2}(t_i, c_k)\}. \tag{18}$$

The pseudo-code of the new algorithm is as follows:

Algorithm 1. Term frequency and distribution based on CHI (TFDCHI).

```
Input:
  D : the training set, C : the category set
  m: the number of selected features
Output:
  F: the top m features in D
Procedure:
1. Init: tf̄_i = 0, tf_{i max} = 0, N_k = 0, sum = 0,
2. TFDCHI1 (t_i, c_k) = 0, TFDCHI2 (t_i, c_k) = 0,
3. TFDCHI1_max(t_i) = 0, TFDCHI2_max(t_i) = 0
4.   for each category c_k ∈ C do
5.     for each document d_j ∈ D do
6.       if d_j ∈ c_k then
7.         N_k ++
8.       end if
9.       for each term t_i ∈ d_j do
10.        sum += tf_ij ;
11.        tf_{i max} = max{tf_{i max}, tf_ij} ; // according to Eq. (13)
12.      end for
13.    end for
14.  for each document d_j ∈ c_k do
15.    for each term t_i ∈ d_j do
16.      tf̄_i = sum/N_k ; // according to Eq. (12)
17.      V(t_i, c_k) += (tf_ij − tf̄_i)^2/(N_k − 1) + α ; // according to Eq. (14)
18.    end for
19.  end for
20.  for each term t_i in class c_k do
21.    TFDCHI1 (t_i, c_k) = log(1 + tf_{i max}) × (t_i, c_k) / V(t_i, c_k) ;
                          // according to Eq. (15)
22.    TFDCHI2 (t_i, c_k) = tf_{i max} × (t_i, c_k) / V(t_i, c_k) ;
                          // according to Eq. (16)
23.    TFDCHI1_max(t_i) = max{TFDCHI1_max(t_i),
                 TFDCHI1(t_i, c_k)} ; // according to Eq. (17)
24.    TFDCHI2_max(t_i) = max{TFDCHI2_max(t_i),
                 TFDCHI2(t_i, c_k)} ; // according to Eq. (18)
```

(*continued*)

25. end for
26. end for
27. F = select_top_features(m); // descending sort $\text{TFDCHI1}_{\max}(t_i)$ and $\text{TFDCHI2}_{\max}(t_i)$ and select m terms
28. return F

A description of the algorithm is given as follows:

Line 1: We initialize some variables which are used in later steps.

Lines 5−13: For each term t_i in the document of category c_k, the total TF is calculated and stored in sum, and we can also obtain the maximum term frequency $tf_{i\,\max}$.

Lines 14−19: By using the value of sum, we can calculate and obtain the average TF. Then, according to Eq. (12), the difference of term distributions is calculated for each term t_i in category c_k.

Lines 20−25: According to Eqs. (15) and (16), the two new approaches calculate the value of term t_i in category c_k . And then, we take the maximum as the value of term t_i in the corpus.

Lines 27−28: We descending sort the corpus' values of all terms and select m terms whose values are maximal as the feature.

4. EXPERIMENTAL SET-UP

4.1 Data-sets

Three data-sets with different characteristics were employed to analyse the behaviour of the proposed method using different types of data:

Reuters-21578: The Reuters corpus contains documents collected from the Reuters newswire in 1987. It is a standard text classification benchmark and contains 135 categories in the original version. Note that Reuters corpus is a much skewed data-set, in which the majority category (*earn*) accounts for 43% of the whole training instances. We adopted a subset of the top 10 categories having 9980 documents. This configuration is also adopted by many previous works [23,40,41]. The stop-words list has 819 words, and then punctuation and numbers are also removed. The total number of unique terms is 58,605.

20-Newsgroup: The 20-Newsgroup corpus is also a widely used benchmark [23,27,37]. It contains 19,997 documents which are taken from the Usenet newsgroup collection, and all documents were assigned evenly to 20 different categories. We randomly select 80% instances from each category as training instances and the rest as test instances. We only keep the information of "Subject" and "Content". Other information, such as "Path", "From", "Message-ID", "newsgroup", "Organization", "References", "Date", and email addresses, are filtered out. The stop-words list has 819 words. The total number of unique terms is 199,704.

WebKB: The WebKB corpus is a collection of 8282 web pages obtained from four academic domains [23,37,40]. The original data-set has seven categories, but only four of them are used: *course*, *faculty*, *project*, and *student*. The total number of unique terms is 21,556.

More details have been showed in Table 3. The "skewness" column of the table shows the skewness property of each data-set, and the last column is the articles which have used the data-sets. Documents are typically represented by VSM. That is to say, the content of a text is represented by a vector in the term space, i.e., $d = \{w_1, w_2, \ldots, w_{|T|}\}$. According to Section 3.2, w_i is the weight of a term t_i, which is calculated by TFIDF (Term Frequency, Inverse Document Frequency) and $|T|$ is the number of terms in the training set.

4.2 Classifiers

In our experiments, we choose kNN [11,40] as the classifier. Besides its simplicity, it is an interesting classifier to evaluate the performance of FS methods because it is strongly influenced by the selected features.

To categorize an unknown document, the kNN classifier ranks the document's neighbours among the training documents and uses the class labels of the k most similar neighbours. Similarity between two documents may be measured by the cosine measure. The similarity score of each nearest neighbour document to the test document is used as the weight of the classes of the neighbour document. If a specific category is shared by more than one of the kNNs, then the sum of the

Table 3: Properties of the data-sets used

Data-set	No. of training documents	No. of test documents	No. of classes	No. of terms	Skewness	Application
Reuters-21578	7193	2787	10	58,605	Highly skewed	[23,40,41]
20-newsgroup	16,000	3997	20	199,704	Homogenous	[23,27,37]
WebKB	2956	1243	4	21,556	Highly skewed	[23,37,40]

similarity scores of those neighbours is obtained from the weight of that particular shared category [38].

At the phase when classification is done by means of the *k*NN, the most important parameter affecting classification is the *k*NN number. Usually, the optimal value of k is empirically determined. In this paper, $k=5$ is determined.

4.3 Performance Measures

For multi-class TC, we measure the effectiveness of the methods using the micro-averaged and macro-averaged F1 [3]. The performance of the F1 classifier for a category is a combination of precision and recall. When effectiveness is computed for several categories, the results for individual categories must be averaged.

For the computation of the micro-averaged F1 (Micro-F1), the category count is proportional to the number positive examples, while in the macro-averaged F1 (Macro-F1), all category counts are considered to be the same. Micro-averaged F1 is dominated by F1 for common categories, while macro-averaged F1 is dominated by F1 for rare categories.

Micro-F1 can be calculated as follows:

$$\text{Micro}-\text{F1}=\frac{2\times R_{\text{micro}}\times P_{\text{micro}}}{(R_{\text{micro}}+P_{\text{micro}})} \tag{19}$$

And the definitions of micro-precision and micro-recall are as follows:

$$P_{\text{micro}}=\frac{\sum_{k}^{|C|}TP_k}{\left(\sum_{k}^{|C|}TP_k+\sum_{k}^{|C|}FN_k\right)}, \tag{20}$$

$$R_{\text{micro}}=\frac{\sum_{k}^{C}TP_k}{\left(\sum_{k}^{|C|}TP_k+\sum_{k}^{|C|}FP_k\right)}, \tag{21}$$

where $|C|$ is the number of categories, TP_k is the number of correctly classified documents for category c_k, FP_k is the number of incorrectly classified documents for category c_k, and FN_k is the number of incorrectly classified documents to other categories else c_k.

Macro-F1 is defined as follows:

$$\text{Macro}-\text{F1}=\frac{2\times R_{\text{macro}}\times P_{\text{macro}}}{(R_{\text{macro}}+P_{\text{macro}})} \tag{22}$$

And macro-precision is calculated as follows:

$$P_{\text{macro}}=\frac{\sum_{k}^{|C|}P_k}{|C|}, \tag{23}$$

$$P_k=\frac{TP_k}{(TP_k+FN_k)}. \tag{24}$$

Micro-recall is calculated as follows:

$$R_{\text{macro}}=\frac{\sum_{k}^{|C|}R_k}{|C|}, \tag{25}$$

$$R_k=\frac{TP_k}{(TP_k+FP_k)}, \tag{26}$$

Where P_k and R_k are the values for a single category c_k.

5. RESULTS

In this section, first, we show the experimental results of our new approach and the classic methods on three corpora with *k*NN classifier. The number of feature spaces is 100, 200, 500, 1000, 2000, 3000, 4000, 5000, 10,000, and 15,000, respectively. And then, we give some discussion for the experimental results.

5.1 Experimental Results

5.1.1 Experimental Results on the Reuters-21578 Data-set

Tables 4 and 5 show the Micro-F1 and the Macro-F1 measure results when *k*NN is used on Reuters-21578 data-set, respectively. Table 4 indicates that the Micro-F1 performance of *k*NN which is based on our proposed approaches is superior to that based on the classic methods except that the number of the selected features is 4000 and 5000. The Micro-F1 of *k*NN based on TFSV2-CHI is the highest (83.03%) among all kinds of feature number. It can be seen from Table 5 that the Macro-F1 performance of *k*NN which is based on our proposed approaches is superior to that based on the classic methods except that the number of the selected features is 500, 3000, and 4000. The Macro-F1 of *k*NN based on TFSV2-CHI is the highest (64.34%) among all kinds of feature number.

5.1.2 Experimental Results on the 20-newsgroups Data-set

Tables 6 and 7 show the Micro-F1 and Macro-F1 measure results when *k*NN is used on 20-newsgroups data-set, respectively. Table 6 indicates that the Micro-F1 performance of *k*NN which is based on our proposed approaches is superior to that based on the classic

Table 4: The Micro-F1 measure results by using *k*NN classifier on Reuters-21578

No. of features	CE (%)	DF (%)	IG (%)	MI (%)	X2 (%)	TFSV1 CHI (%)	TFSV2 CHI (%)
100	79.94	75.78	79.73	53.93	79.26	57.81	**79.94**
200	81.84	77.93	81.23	54.43	80.77	58.34	**83.03**
500	82.60	80.66	82.63	60.78	82.78	61.77	**82.99**
1000	81.31	80.95	81.13	66.52	80.91	64.21	**81.59**
2000	81.02	81.20	80.91	71.55	80.84	76.59	**81.88**
3000	80.80	81.23	80.73	72.87	80.77	78.19	**81.23**
4000	80.37	**80.88**	80.37	74.31	80.34	79.27	80.09
5000	80.23	**80.62**	80.23	74.81	79.94	80.42	79.91
10,000	79.98	80.12	79.98	78.94	79.87	**81.39**	79.12
15,000	79.81	80.09	79.80	79.89	79.78	**81.95**	78.36

Note: The bold values indicate the best performance of the classifier when various feature selection methods are used respectively.

Table 5: The Macro-F1 measure results by using *k*NN classifier on Reuters-21578

No. of features	CE (%)	DF (%)	IG (%)	MI (%)	X2 (%)	TFSV1 CHI (%)	TFSV2 CHI (%)
100	61.01	50.76	60.51	26.21	59.48	43.68	61.77
200	62.81	53.51	62.06	26.23	61.11	46.77	64.34
500	64.00	60.32	63.76	34.31	64.09	49.50	62.72
1000	62.34	60.97	59.62	39.93	62.32	58.01	62.36
2000	59.86	59.47	59.70	45.40	61.80	60.23	63.16
3000	59.07	**62.63**	59.31	49.99	61.91	60.98	60.21
4000	60.98	**61.63**	60.99	51.47	57.63	61.22	58.87
5000	60.70	59.10	60.77	51.41	60.56	**61.34**	60.79
10,000	60.95	61.40	60.95	57.29	61.03	**61.56**	59.48
15,000	61.22	61.70	61.35	60.55	61.29	**62.03**	60.18

Note: The bold values indicate the best performance of the classifier when various feature selection methods are used respectively.

methods except that the number of the selected features is 1000 and 2000. The Micro-F1 of *k*NN based on TFSV2-CHI is the highest (65.05%) among all kinds of feature number. As shown in Table 7, the Macro-F1 performance of *k*NN which is based on our proposed approaches is superior to that based on the classic methods except that the number of the selected features is 1000.

The Macro-F1 of *k*NN based on TFSV2-CHI is the highest (65.81%) among all kinds of feature number.

5.1.3 *Experimental Results on the WebKB Data-set*

Tables 8 and 9 show the Micro-F1 and Macro-F1 measure results when *k*NN is used on WebKB, respectively. Table 8 indicates that the Micro-F1 performance of *k*NN

Table 6: The Micro-F1 measure results by using *k*NN classifier on 20-newsgroup

No. of features	CE (%)	DF (%)	IG (%)	MI (%)	X2 (%)	TFSV1 CHI (%)	TFSV2 CHI (%)
100	54.20	36.40	53.50	20.50	55.65	51.33	**57.60**
200	56.30	43.55	52.85	23.80	57.30	51.96	**62.40**
500	57.35	46.80	52.90	29.50	55.40	53.28	**65.05**
1000	**59.35**	48.15	53.45	33.50	55.40	54.17	56.20
2000	**57.10**	52.95	56.90	37.60	56.15	55.52	57.05
3000	57.15	55.75	57.05	42.05	56.40	56.81	**58.60**
4000	57.75	55.75	57.60	42.05	56.20	57.23	**59.75**
5000	57.90	56.30	57.30	45.45	58.25	58.64	**59.35**
10,000	57.05	57.75	56.95	51.15	58.25	**59.13**	58.40
15,000	56.80	57.50	56.86	56.12	57.95	**59.87**	58.01

Note: The bold values indicate the best performance of the classifier when various feature selection methods are used respectively.

Table 7: The Macro-F1 measure results by using *k*NN classifier on 20-newsgroup

No. of features	CE (%)	DF (%)	IG (%)	MI (%)	X2 (%)	TFSV1 CHI (%)	TFSV2 CHI (%)
100	57.34	37.69	56.24	20.51	60.15	53.18	**63.39**
200	56.95	44.35	53.86	23.61	58.93	53.89	**65.31**
500	58.13	48.70	54.20	29.59	56.68	55.16	**65.81**
1000	**59.58**	50.12	55.04	33.97	56.45	56.01	57.34
2000	58.02	54.66	57.89	38.92	57.24	56.75	**58.40**
3000	58.72	57.05	58.54	43.96	57.68	57.44	**59.75**
4000	59.08	57.04	58.79	44.17	57.63	58.61	**61.02**
5000	59.13	57.38	58.47	47.48	59.42	59.38	**61.02**
10,000	58.72	59.03	58.64	52.89	59.59	**60.07**	59.69
15,000	58.56	58.54	58.60	56.36	59.23	**60.59**	59.60

Note: The bold values indicate the best performance of the classifier when various feature selection methods are used respectively.

Table 8: The Micro-F1 measure results by using *k*NN classifier on WebKB

No. of features	CE (%)	DF (%)	IG (%)	MI (%)	X2 (%)	TFSV1 CHI (%)	TFSV2 CHI (%)
100	64.39	56.45	62.18	46.51	63.10	60.21	**68.66**
200	68.11	60.02	69.46	49.12	70.32	63.15	**72.23**
500	73.98	67.45	74.84	53.24	75.09	68.39	**77.00**
1000	78.10	72.88	77.70	56.50	77.63	77.02	**78.10**
2000	78.15	76.14	78.10	60.62	78.27	78.37	**79.81**
3000	78.65	78.00	79.31	65.95	79.58	79.56	**81.06**
4000	79.61	78.80	80.01	67.60	80.39	80.63	**81.62**
5000	80.21	79.71	80.96	69.36	80.94	81.44	**81.47**
10,000	80.88	80.71	81.23	71.11	81.10	**81.93**	80.86
15,000	81.19	81.08	81.56	72.99	81.33	**82.28**	80.48

Note: The bold values indicate the best performance of the classifier when various feature selection methods are used respectively.

Table 9: The Macro-F1 measure results by using *k*NN classifier on WebKB

No. of features	CE (%)	DF (%)	IG (%)	MI (%)	X2 (%)	TFSV1 CHI (%)	TFSV2 CHI (%)
100	55.94	48.80	54.78	38.20	55.01	51.42	**56.82**
200	59.57	53.64	60.52	41.46	59.60	52.55	**60.90**
500	64.83	59.62	64.82	46.02	64.65	57.01	**66.11**
1000	67.94	64.36	68.17	50.62	67.79	66.56	**68.32**
2000	68.79	66.78	68.75	55.05	68.47	68.47	**69.99**
3000	68.83	68.69	69.75	60.01	69.62	69.56	**71.41**
4000	69.83	69.36	70.25	61.00	70.38	70.41	**71.73**
5000	70.58	70.14	71.09	61.84	70.99	71.13	**71.49**
10,000	71.09	70.83	71.54	63.52	71.48	**71.59**	71.01
15,000	71.32	71.02	71.89	65.23	71.69	**71.89**	70.58

Note: The bold values indicate the best performance of the classifier when various feature selection methods are used respectively.

which is based on our proposed approaches is superior to classic methods. The Micro-F1 of *k*NN based on TFSV1-CHI is the highest (82.28%). As shown in Table 9, the Macro-F1 performance of *k*NN which is based on our proposed approaches is superior to that based on the classic methods. The Macro-F1 of *k*NN based on TFSV2-CHI is the highest (71.89%).

5.2 Discussion

Our proposed approaches are based on TF and the term distribution, which can overcome "the drawback of low-frequency terms and the neglect of the term distribution" of classic CHI FS methods. Through extensive experiments on three common text corpora with *k*NN

classifier, we obtain some exciting results. Now we give some discussion as follows:

Because of the weakness of low-frequency terms of CHI, we give a modified parameter which is TF. Then, we consider the basic theory of chi-square in Section 3.1, the greater the value of CHI, the more category information the feature owns; therefore, we get the maximum TF into our new approaches. According to Eq. (15) and Eq. (16), this modified parameter makes the value of the term whose DF is low and TF is high to be large. Another modified parameter which is the term distribution is also important. According to the definition in Section 3.2, we want the sample variance of the selected features smaller, and then the value which is calculated by Eqs. (15) and (16) is increasing further. That is to say, some discerning terms whose TF is low can be chosen as features.

Through extensive experiments on three common text corpora with *k*NN classifier, we observe that if TFDCHI1 is used to select feature, the Micro-F1 and Macro-F1 performances are slightly better than other methods only in the situation that we have a lot of features (10,000 and 15,000). Therefore, we think deeply about this situation and propose TFDCHI2. Then, we can obtain the best Micro-F1 and Macro-F1 performances, when the number of selected features is few. If we want to use less features to accomplish TC effectively, the second approach is the best.

6. CONCLUSION AND FUTURE WORK

TF often impacts the topics of corpus; therefore, we propose modified approaches based on CHI. In these approaches, we also consider the term distribution which is another important parameter for CHI. Through extensive experiments on three common text corpora with *k*NN classifier, we can observe that the Micro-F1 and Macro-F1 performances of our approach are better than the classic methods in most instances. In future work, we will research some other parameters which can improve the classic method, and we also will find some way to optimize the time complexity of our approaches.

Funding

This work was supported in part by National Science Foundation of China [grant number 61173143], [grant number 61100007], [grant number 61100081]; China Postdoctoral Science Foundation [grant number 2012M511303], and was also supported by PAPD (A Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions).

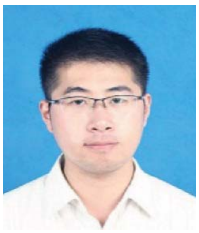
The authors extend their appreciation to the Deanship of Scientific Research at King Saud University for funding this work through research [grant number RGP-264].

REFERENCES

1. M. G. H. AlZamil, and A. B. Can, "ROLEX-SP: rules of lexical syntactic patterns for free text categorization," *Knowl. Based Syst*., Vol. 24 pp. 58–65, Feb. 2011.
2. F. Sebastiani, "A tutorial on automated text categorisation," in *Proceedings of the ASAI-99 in 1st Argentinian Symposium on Artificial Intelligence*, Buenos Aires, AR, 1999, pp. 17–35.
3. F. Sebastiani, "Machine learning in automated text categorization," *ACM Comput. Surv.*, Vol. 34, no. 1, pp. 1–47, Jan. 2002.
4. W. Shang, H. Huang, H. Zhu, Y. Lin, Y. Qu, and Z. Wang, "A novel feature selection algorithm for text categorization," *Expert Syst. Appl*., Vol. 33, no. 1, pp. 1–5, Jan. 2007.
5. K. Coussement, and D. Van den Poel, "Integrating the voice of customers through call center emails into a decision support system for churn prediction," *Inf. Manag*., Vol. 45, no. 3, pp. 164–74, Mar. 2008.
6. G. Sakkis, I. Androutsopoulos, G. Paliouras, V. Karkaletsis, C. D. Spyropoulos, and P. Stamatopoulos, "A memory-based approach to anti-spam filtering for mailing lists," *Inf. Retr.*, Vol. 6, no. 1, pp. 49–73, Jan. 2003.
7. B. Zhou, Y. Y. Yao, and J. Luo, "A three-way decision approach to e-mail spam filtering," in *Proceedings of 23rd Canadian Conference on Artificial Intelligence (Canadian AI'10)*, Ottawa, Vol. 6085, Lecture Notes in Computer Science, 2010, pp. 28–39.
8. X. Qi, and B. D. Davison, "Web page classification: features and algorithms," *ACM Comput. Surv.*, Vol. 41, no. 2, pp. 1–31, Feb. 2009.
9. S. Wang, D. Li, X. Song, Y. Wei, and H. Li. "A feature selection method based on improved Fisher's discriminant ratio for text sentiment classification," *Expert Syst. Appl*., Vol. 38, no. 7, pp. 8696–702, Jul. 2011.
10. D. Wang, H. Zhang, R. Liu, M. Lin, and W. Wu, "Predicting bugs components via mining bug reports," *J. Softw.*, Vol. 7, no. 5, pp. 1149–54, May 2012.
11. Y. Yang, and J. O. Pedersen, "A comparative study on feature selection in text categorization," in *Proceedings of the 14th International Conference on Machine Learning*, Nashville, TN, 1997, pp. 412–20.
12. C. Cortes, and V. Vapnik, "Support-vector networks," *Mach. Learn*., Vol. 20, no. 3, pp. 273–97, Mar. 1995.
13. D. Lewis, "Naive (Bayes) at forty: the independence assumption in information retrieval," in *European Conference on Machine Learning*, Chemnitz, 1998, pp. 4–15.
14. A. McCallum, and K. Nigam, "A comparison of event models for Naive Bayes text classification." in *Workshop on Learning for Text Categorization*, San Francisco, CA, 1998, pp. 41–8.
15. J. Zhang, L. Chen, and G. Guo, "Projected-prototype based classifier for text categorization," *Knowl. Based Syst*., Vol. 49, pp. 179–89, Sep. 2013.
16. T. T. Nguyen, K. Chang, and S. C. Hui, "Supervised term weighting centroid-based classifiers for text categorization," *Knowl. Inf. Syst*., Vol. 35, no. 1, pp. 61–85, Apr. 2013.
17. I. Guyon, and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res*., Vol. 3, pp. 1157–82, Mar. 2003.
18. D. Wang, H. Zhang, R. Liu, and W. Lv, "Feature selection based on term frequency and *T*-test for text categorization," in *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, Maui, 2012, pp. 1482–6.
19. N. Azam, and J. Yao, "Comparison of term frequency and document frequency based feature selection metrics in text categorization," *Expert Syst. Appl*., Vol. 39, pp. 4760–8, May 2012.
20. Y. Zheng, X. Cheng, R. Huang, and Y. Man, "A comparative study on unsupervised feature selection methods for text clustering," in *Proceeding of Second International Conference, ADMA 2006*, Xi'an, 2006, pp. 644–51.
21. A. L. Blum, and P. Langley, "Selection of relevant features and examples in machine learning," *Artif. Intell*., Vol. 97, pp. 245–71, Jan. 1997.

22. D. Mladenic, and M. Grobelnik, "Feature selection on hierarchy of web documents," *Decis. Support Syst*., Vol. 35, pp. 45–87, Jan. 2003.
23. J. Yang, Y. Liu, X. Zhu, Z. Liu, and X. Zhang, "A new feature selection based on comprehensive measurement both in inter-category and intra-category for text categorization," *Inf. Process. Manag*., Vol. 48, pp. 741–54, Jul. 2012.
24. S. Tasc, and T. Güngör, "Comparison of text feature selection policies and using an adaptive framework," *Expert Syst. Appl*., Vol. 40, pp. 4871–86, Jul. 2013.
25. H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans. Pattern Anal. Mach. Intell*., Vol. 27, pp. 1226–38, Aug. 2005.
26. D. Koller, and M. Sahami, "Hierarchically classifying documents using very few words," in *Proceedings of ICML*, Nashville, TN, 1997, pp. 170–8.
27. D. Wang, H. Zhang, R. Liu, W. Lv, and D. Wang, "t-Test feature selection approach based on term frequency for text Categorization," *Pattern Recogn. Lett*., Vol. 45, pp. 1–10, Aug. 2014.
28. L. Dai, J. Hu, and W. Liu, "Using modified CHI square and rough set for text categorization with many redundant features," in *International Symposium on Computational Intelligence and Design*, Wuhan, China, Oct. 2008, pp. 182–185.
29. R. M. Yunus, and S. Khan, "The bivariate noncentral chi-square distribution – a compound distribution approach," *Appl. Math. Comput*., Vol. 217, pp. 6237–47, Dec. 2011.
30. T. Xia, Y. Chai, H. Lu, and T. Wang, "An improved global weight function of terms based on Pearson's Chi-square," in *Information Science and Control Engineering (ICISCE)*, Shenzhen, 2012, pp. 1–5.
31. G. Evgeniy, and M. Shaul, "Text categorization with many redundant features: using aggressive feature selection to make SVMs competitive with C4.5," in *The 21st International Conference on Machine Learning (ICML)*, Banff, 2004, pp. 321–8.
32. P. Cerchiello, and P. Giudici, "Non parametric statistical models for on-line text classification," *Adv. Data Anal. Classif*., Vol. 6, pp. 277–88, Apr. 2012.
33. Y.-T. Chen, and M. C. Chen, "Using chi-square statistics to measure similarities for text categorization," *Expert Syst. Appl*., Vol. 38, pp. 3085–90, Apr. 2011.
34. L.-H. Lee, and C.-J. Luh, "Generation of pornographic blacklist and its incremental update using an inverse chi-square based method," *Inform. Process. Manag*., Vol. 44, pp. 1698–706, Oct. 2008.
35. J. Tian, D. Yu, B. Yu, and S. Ma, "A fuzzy TOPSIS model via chi-square test for information source selection," *Knowl. Based Syst*., Vol. 37, pp. 515–27, Jan. 2013.
36. J. Lee, and D.-W. Kim, "Feature selection for multi-label classification using multivariate mutual information," *Pattern Recogn. Lett*., Vol. 34, pp. 349–57, Feb. 2013.
37. C. Shang, M. Li, S. Feng, Q. Jiang, and J. Fan, "Feature selection via maximizing global information gain for text classification," *Knowl. Based Syst*., Vol. 54, pp. 298–309, Oct. 2013.
38. H. Uguz, "A two-stage feature selection method for text categorization by using information gain, principal component analysis and genetic algorithm," *Knowl. Based Syst*., Vol. 24, pp. 1024–32, Oct. 2011.
39. X. Sheng, and M. Jiang, "Automatic classification of Chinese documents based on rough set and improved quick-reduce algorithm," *J. Electron. Inf. Technol*., Vol. 27, no. 7, pp. 1047–52, Jul. 2005.
40. R. H. W. Pinheiro, G. D. C. Cavalcanti, R. F. Correa, and T. I. Ren, "A global-ranking local feature selection method for text categorization," *Expert Syst. Appl*., Vol. 39, pp. 12851–7, Dec. 2012.
41. Y. Chang, S. Chen, and C. Liau, "Multilabel text categorization based on a new linear classifier learning method and a category-sensitive refinement method," *Expert Syst. Appl*., Vol. 34, pp. 1948–53, Apr. 2008.

Authors



Chuanxin Jin received his bachelor's degree in Computer Science and Engineering from Nanjing University of Information Science & Technology, China, in 2009. Currently, he is a candidate for the degree of Master of Computer Science and Engineering in Nanjing University of Information Science & Technology. His research interests include cloud computing, parallel computing and data mining, etc.

E-mail: jcxzl1018@163.com

Tinghuai Ma is a professor in Computer Sciences at Nanjing University of Information Science & Technology, China. He received his bachelor (HUST, China, 1997), master (HUST, China, 2000), PhD (Chinese Academy of Science, 2003) degrees, and was a post-doctoral associate (AJOU University, 2004). From November 2007 to July 2008, he visited Chinese Meteorology Administration. From February 2009 to August 2009, he was a visiting professor in Ubiquitous computing Lab, Kyung Hee University. His research interests are data mining, cloud computing, ubiquitous computing, privacy preserving, etc. He has published more than 100 journal/conference papers.

E-mail: thma@nuist.edu.cn

Rongtao Hou is a professor in Computer Sciences at Nanjing University of Information Science & Technology, China. He received his PhD degree from China North-East University in 2003. His research interests are mobile ad-hoc networks, wireless sensor networks, internet of things, and meteorological observation systems.

E-mail: rthou@tom.com

Meili Tang is an associate professor in School of Public Administration at Nanjing University of Information Science & Technology, China. She received her master's degree from Huazhong University of Science & Technology, China, in 2000. Her main research interests are in the areas of e-government and data publishing.

E-mail:meilitg@263.com

Yuan Tian has received her master's and PhD degrees from KyungHee University, and she is currently working as an assistant professor at College of Computer and Information Sciences, King Saud University, Kingdom of Saudi Arabia. She is a member of technical committees of several international conferences. In addition, she is an active reviewer of many international journals. Her research interests are broadly divided into privacy and security, which are related to cloud computing, bioinformatics, multimedia, cryptograph, smart environment, and big data.

E-mail: ytian@ksu.edu.sa

Abdullah Al-Dhelaan has received his BS degree in Statistics (Hon) from King Saud University on 1982, and the MS and PhD degrees in Computer Science from Oregon State University on 1986 and 1989, respectively. He is currently the vice dean for Academic Affairs, deanship of Graduate Studies, and a professor of Computer Science, King Saud University, Riyadh, Saudi Arabia. He has guest edited several special issues for the *Telecommunication Journal* (Springer) and *the International Journal for Computers and Their Applications* (ISCA). Moreover, he is currently on the editorial boards of several journals and the organizing committees for several reputable international conferences. His current research interest includes mobile ad hoc networks, sensor networks, cognitive networks, network security, image processing, and high-performance computing.

E-mail: dhelaan @ksu.edu.sa

Mznah Al-Rodhaan has received her BS degree in Computer Applications (Hon) and MS degree in Computer Science both from King Saud University on 1999 and 2003, respectively. In 2009, she received her PhD degree in Computer Science from University of Glasgow in Scotland, UK. She is currently working as the vice chair of the Computer Science Department in College of Computer & Information Sciences, King Saud University, Riyadh, Saudi Arabia. Moreover, she has served in the editorial boards for some journals such as the *Ad Hoc Journal* (Elsevier), and has participated in several international conferences. Her current research interest includes mobile ad hoc networks, wireless sensor networks, multimedia sensor networks, cognitive networks, and network security.

E-mail: rodhaan@ksu.edu.sa

DOI: 10.1080/03772063.2015.1021385;