PAPER

Improving Automatic Text Classification by Integrated Feature Analysis

Lazaro S.P. BUSAGALA[†a)], ***Nonmember***, **Wataru OHYAMA**[†b)], **Tetsushi WAKABAYASHI**[†c)], ***and*** **Fumitaka KIMURA**[†d)], ***Members***

SUMMARY Feature transformation in automatic text classification (ATC) can lead to better classification performance. Furthermore dimensionality reduction is important in ATC. Hence, feature transformation and dimensionality reduction are performed to obtain lower computational costs with improved classification performance. However, feature transformation and dimension reduction techniques have been conventionally considered in isolation. In such cases classification performance can be lower than when integrated. Therefore, we propose an integrated feature analysis approach which improves the classification performance at lower dimensionality. Moreover, we propose a multiple feature integration technique which also improves classification effectiveness.

key words: *text classification/categorization, feature transformation, dimension reduction, principal component analysis, canonical discriminant analysis, integrated feature analysis, multiple feature integration*

1. Introduction

Automatic text classification (ATC) is a task that involves high dimensional space, which needs feature selection and reduction. Feature extraction and reduction are also important research areas in pattern recognition and in other related fields. The main advantages of feature selection and reduction include the use of smaller amounts of features, hence lower computational complexity. Other advantages include the use of fewer features in relation to sample size, leading to more accurate density estimation and higher classification performance.

In ATC, the high dimensionality problem is the result of the increase in the number of words. This increase goes along with increase in the dimensionality to use in the classification process. For instance, in this work the dimensionality of feature vectors in all utilized articles from the Reuters-21578 corpus amounted to 24,868 words. Such a high dimensional feature space needs large amounts of calculation resources and memory storage capacity for processing and classification.

In order to solve this problem, dimension reduction is required. Preferably, the process of dimension reduction should not lower the classification performance. In ATC, various dimension reduction techniques are used. Conventionally, most are used in isolation. Frequently, the dimensionality reduction technique that has been conventionally used is the latent semantic indexing (LSI) [1]. In recent years, some works have applied principal component analysis (PCA) to reduce the dimensionality [2].

However, it has been noted that PCA [3] and LSI [4] ignore category specific information. Since PCA maximizes the total scatter across all classes, it can result in retention of non-discriminative information. To acquire more discriminative information, it can be desirable to apply canonical discriminant analysis (CDA) to fewer extracted components. Nevertheless, the direct application of CDA can lead to a singularity problem of the within-class scatter matrix. Meanwhile transformed features in ATC can yield better classification performance. It is therefore desirable to perform feature transformation to improve class document separability.

We therefore propose an approach called integrated feature analysis (IFA) which includes the normalization of absolute word frequency to relative frequency and power transformation, PCA and CDA in an integrated combination approach.

Moreover, we propose multiple features integration (MFI) to improve class document separability. We integrate features at the lowest dimensionality possible. In so doing, we improve the classifier's efficiency while improving classification effectiveness.

The rest of this paper is organized as follows. Section 2 describes the proposed methodological framework. Section 3 explains some implementation issues including the classification experiments to verify the effectiveness of the proposed framework. In Sect. 4, we discuss the experimental results. Section 5 gives a survey of related works. Finally, a summary and future research possibilities are given in Sect. 6.

2. Integrated Feature Analysis (IFA)

This section describes techniques on which the proposed approach is based. We name this framework the integrated feature analysis (IFA). This includes the descriptions of the multiple feature integration (MFI).

2.1 General Steps of IFA

The IFA framework is based on the idea of integrating var-

Manuscript received June 19, 2007.
Manuscript revised November 7, 2007.

[†] The authors are with the Graduate School of Engineering, Mie University, Tsu-shi, 514–8507, Japan.
a) E-mail: busagala@hi.info.mie-u.ac.jp
b) E-mail: ohyama@hi.info.mie-u.ac.jp
c) E-mail: waka@hi.info.mie-u.ac.jp
d) E-mail: kimura@hi.info.mie-u.ac.jp

DOI: 10.1093/ietisy/e91–d.4.1101

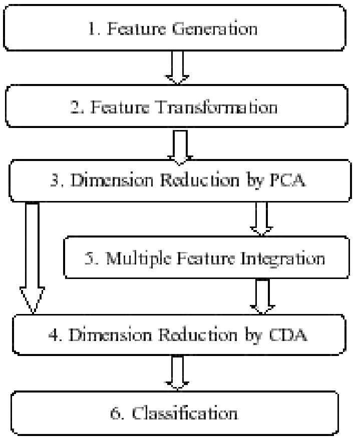

1. Feature Generation
2. Feature Transformation
3. Dimension Reduction by PCA
5. Multiple Feature Integration
4. Dimension Reduction by CDA
6. Classification

Fig. 1 General steps of integrated feature analysis.

ious techniques such as feature transformation, dimension reduction techniques and the integration of multiple features to generate composite features. By considering the techniques together, we achieve a better optimization of class separability than when they are considered in isolation. We provide an overview of the feature transformations, principal components analysis (PCA) and canonical discriminant analysis (CDA).

We then introduce a novel technique in this domain, which we call multiple feature integration (MFI). After integrating all the other techniques, we use CDA on fewer extracted features. An overview of the proposed approach is given in Fig. 1, which shows the general steps of the IFA. We discuss step 1 and 2 in Sect. 2.2.

The aim of applying PCA in step 3 is to solve the problem of singularity of the within-class scatter matrix, which can result from a smaller sample size than the vocabulary list. Step 4 uses fewer dominant components obtained from step 3. Therefore, while extracting fewer features for the classification system, the document class separability and the computation efficiency can be improved.

Note that, step 3 and 4 cannot be interchanged because CDA is vulnerable to the singularity problem. On the other hand, PCA has a limitation of retaining non-discriminative information. Thus, interchanging them may lower classification performance. Also, it is noteworthy that instead of PCA other dimension reduction techniques that do not suffer from the singularity problem can be plugged in. A technique such as latent semantic indexing (LSI) can be a good candidate to use. Similarly class discriminative techniques such as nonparametric discriminant analysis can be used instead of CDA. In the following subsections we discuss our approach in more detail.

2.2 Feature Vector Generation and Transformation

The first step is to generate feature vectors to represent the textual documents. Let us consider a set of N sample texts, $\chi = \{X_1, X_2, \cdots, X_N\}$ with n–dimensional text space. In addition, let us assume that every textual document belongs to one of the C classes $\{\omega_1, \omega_2, \ldots, \omega_C\}$. Each text can be represented as a feature vector X which can be denoted as

$$X = [x_1 \;\; x_2 \ldots x_n]^T. \tag{1}$$

Where, n is dimensionality (lexicon size), x_i is the frequency value of i^{th} word and T refers to the transpose of a vector. It is clear that the feature vectors generated in this way are absolute term frequencies.

Feature transformation in step 2 of Fig. 1, refers to normalization of absolute term frequency to relative frequency and power transformation. These are defined in this subsection. In our approach, feature transformations are performed to minimize problems that can be encountered during classification. Specifically, we transform absolute word frequency to relative word frequency (RF) to solve the problem of dependency on text length as follows.

$$y_i = \frac{x_i}{\sum_{j=1}^{n} x_j}. \tag{2}$$

It follows that $\sum_{i=1}^{n} y_i = 1$. Therefore, the problem of dependency on text length within a class is solved.

After obtaining RF, the document sample distribution may still be skewed. This may lead to generation of classification errors. This is even more problematic for classifiers such as linear or quadratic classifiers, which are typically designed for Gaussian distributions. Hence, with the purpose of obtaining Gaussian-like distribution, power transformation is performed as follows.

$$z_i = y_i^v \; (0 < v < 1). \tag{3}$$

This transformation generates Gaussian-like sample distribution, leading to better classification performance. Further details of the power transformation technique are found in [5].

2.3 Dimension Reduction by Principal Component Analysis (PCA)

As noted in Sect. 1, the problem of high dimensional space after feature transformations remains. To solve the problem we apply principal component analysis [3]. For the convenience of the reader, a short review is given as follows. From the set of training documents $\chi = \{X_1, X_2, \cdots, X_N\}$ the total covariance matrix Σ of the training sample is computed as

$$\Sigma = \frac{1}{N} \sum_{X \in \chi} (X - M)(X - M)^T; \tag{4}$$

$$M = \frac{1}{N} \sum_{X \in \chi} X. \tag{5}$$

Where, M is the total mean vector of the training sample.

Corresponding eigenvalues λ_i and eigenvectors Φ_i are obtained by the definition:

$$\Sigma \Phi_i = \lambda_i \Phi_i (i = 1, 2, \ldots, n), \tag{6}$$

provided that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$. Using eigenvectors corresponding to the $m(m \leq n)$ largest eigenvalues, principal components z_i are defined by the linear transformation

$$z_i = \Phi_i^T X (i = 1, 2, \cdots, m). \tag{7}$$

The reduced dimension of feature vectors can be obtained from m principal components selected to compose m-dimensional feature space.

2.4 Dimension Reduction by Canonical Discriminant Analysis (CDA)

One of the drawbacks of PCA is to ignore category specific information. Because, it maximizes the total scatter across all classes (i.e. total covariance matrix), resulting in retention of non-discriminative information.

To avoid rank deficiency and singularity problems after reducing the dimensionality using PCA, we use the appropriate amount of features to which canonical discriminant analysis (CDA) is applied. CDA considers category specific information by using the between-class and the within-class scatter matrices that generalize Eq. (6) as defined by

$$S_B \Phi_i = \lambda_i S_W \Phi_i, \tag{8}$$

where S_B and S_W are between-class and within-class covariance matrices, respectively. Their definitions follow such that

$$S_B = \sum_{j=1}^{C} \frac{N_j}{N} (M_j - M)(M_j - M)^T, \tag{9}$$

and

$$S_W = \frac{1}{N} \sum_{j=1}^{C} \sum_{X \in \chi_j} (X - M_j)(X - M_j)^T, \tag{10}$$

where the mean vector for each class M_j is defined by

$$M_j = \frac{1}{N_j} \sum_{X \in \chi_j} X. \tag{11}$$

N_j and χ_j refer to the number of documents and the set of text samples in a particular class w_j respectively. The other symbols mean the same as defined before.

The canonical discriminants z_i can be obtained using Eq. (7). As we assumed that there are C classes in the data collection, it is noteworthy that since S_B results from the sum of C matrices and each matrix is of rank 1 or less, then S_B is of rank $C - 1$. Therefore, there are at most $C - 1$ nonzero generalized eigenvalues and their vectors. This implies that $C - 1$ dimensional space or less might give highest classification performance.

2.5 Multiple Feature Integration (MFI)

Multiple feature integration (MFI) refers to the integration of multiple features to generate composite features that give higher classification effectiveness. The objective is to improve the separability between class documents by integrating features using the concept from set theory commonly known as the union of sets. By suitably integrating reduced transformed and reduced term weighted features, we can achieve improved separability of class documents.

As noted in Sect. 2.4, the rank of the between-class matrix leads to $C - 1$ dimensional space that gives higher classification performance. This means the dimensionality highly depends on the number of classes available. This might not be satisfactory. It is an unsatisfactory condition because the separability may not be brought to maximum level. In other words, classification effectiveness might not necessarily be improved to 100%. This implies that there is room for improvement. As one way of improving the situation, we propose multiple feature integration (MFI).

More formally, for the purpose of describing this technique let us assume that there are two feature vectors called weighted term frequency by inverse document frequency (TFIDF) and non-weighted term frequency (TF). The TF features may be transformed as described in Sect. 2.2. Furthermore, let us assume that PCA has been performed, and we can denote such feature vectors as

$$X^{tf} = \left[x_1^{tf}, x_2^{tf}, \ldots, x_n^{tf} \right], \tag{12}$$

for the TF features. The TFIDF weighted document features can be expressed as

$$X^{w} = \left[x_1^{w}, x_2^{w}, \cdots, x_n^{w} \right]. \tag{13}$$

Assume that $\oplus$ represents the union of two vectors. The combination of these document features can be expressed as the union of the two as

$$\begin{aligned} CF &= \left[X^{tf} \oplus X^{w} \right] \\ &= \left[x_1^{tf}, x_2^{w}, x_3^{tf}, x_4^{w} \ldots, x_n^{tf}, x_{(n)}^{w} \right]. \end{aligned}$$

Where X^{tf} refers to the non-weighted term frequency features, which can be the absolute term frequency (AF) or the power transformed features (PT) or the relative term frequency (RF) or the relative term frequency followed by power transformation (RFPT). The feature vectors to be selected for the integration should be those achieving the highest classification performances. Similarly, X^w should be from the feature vectors namely TFIDF.

The reason for the improved class document separability is that the integrated composite feature CF combines the discriminative information from both TF and TFIDF. Specifically, RFPT is advantageous, because it has no dependency on document length and its sample distribution has the properties of Gaussian-like distribution (recall from Sect. 2.2). The TFIDF discriminative information lies in the fact that it assigns a high degree of importance to terms that occur in only a few documents of the text data set. We use this technique i.e., MFI to generate composite features to improve classification effectiveness. We verify the impact of MFI in Sect. 4.2.

2.6 Advantages of IFA

We state various advantages in this Subsection. These can also be regarded as the reasons for the improved classification effectiveness.

There are various advantages for IFA. First, IFA uses transformed features which do not depend on textual document length; hence, within-class variability can be avoided. Second, as it is described in Sect. 2.2, transformed features provide a Gaussian-like sample distribution which improves the separability of documents by the classification system. Third, the singularity problem emanating from smaller sample size than its dimensionality is solved at the PCA stage.

The fourth advantage can be realized by improving the separability by applying CDA. This is applied to the dominant principal components rather than directly to the original features. Hence, numerical stableness and classifier's efficiency can be realized. The fifth advantage is brought by MFI. This improves further the class document separability. Moreover, MFI can incorporate new features other than from TFIDF and RFPT. The end result is improved classification performance. In short, it can be said that the limitations of each technique can be avoided by using IFA.

3. Experiments

This section gives an account on the implementation issues. Specifically, we describe the data for experiments, the adopted feature selection methods, the techniques used for dimension reduction, the classification process and the performance measures that we applied.

3.1 Data for Experiments

For effective evaluation of the proposed methods, a set of text data for category assignment is required. A benchmark collection for text categorization research called Reuters-21578 was used. This collection has been widely employed by other researchers too [6]–[12]. Reuters-21578 is composed of 21,578 articles, which are manually classified into 135 categories, and one textual document belongs to one or more categories. Hence it is both a multi-class and multi-label problem.

We used the ModApte Split, which contains 12,902 articles. In this split, the training set contains 9,603 documents. The test set contains 3,299 documents, and 8,676 documents are not used. ModApte Split is the most commonly used. In total, we used 115 categories in the experiments. We also use a 90 category set for further comparability. We report results of 90 categories in Table 1 only.

3.2 Lexicon Generation

In general, function words are not useful to represent document features discriminatingly. Therefore, before generating feature vectors, functional words and general words were removed with reference to a stop list prepared beforehand. This process reduces the features for the classification systems. This also reduces the amount of memory required for storage as well as processing time required by the classification systems.

Even when the stop list was used to remove useless words, many words remained. Hence, words with frequency value of 5 or less in all the training data were removed. The objective was to reduce further the remaining words. This removal of words reduced the lexicon size from 24868 to 7474. According to [1], this word removal does not affect the classification performance.

3.3 Implementation on Dimension Reduction

As described in Sect. 3, PCA was used to reduce the dimensionality from 7474 to 1000 before the application of CDA. The dimensionality of 1000 was chosen based on the experiments which showed no classification performance improvement even when more features were added afterward. We use the feature vectors with 1000 dimensionality in the CDA algorithm. CDA with C classes gives at most $C-1$ nonzero generalized eigenvalues and eigenvectors. We therefore chose the dimensionality in the experiments to be 114, when 115 categories are used. In the 90 category set 89 word features were used.

3.4 Classification and Performance Measures

Various classification methods have been proposed in the literature [1], [2], [8], [9], [13]. Among others, k nearest neighbor (kNN) is one of the best performers. The kNN algorithm rely on the concept that given a test document, the system finds the k nearest neighbors among the training documents to predict its category [13].

Unlike other classification methods, kNN can easily handle both multi-class and multi-label problems simultaneously. Since the Reuters-21578 collection is both a multi-class and multi-label problem, kNN was used in the classification process. To determine the kNN, cosine similarity function was used in the experiments. The k value was experimentally varied from 1 until when the classifier could give more errors rather than improving the classification performance. The results reported in this paper are the highest in respective type of features.

We adopt the recall, precision and F-measure for the classification performance evaluation. These measures are regarded as standard evaluation methods for the classification systems in automatic text classification. The definitions of these measures can be found in [1], [13]. Micro-averaging and macro-averaging strategies are usually adopted. For comparability with other previous works in the literature, we adopt micro-averaging strategy too.

4. Empirical Results

This section discusses the results of the experiments. For

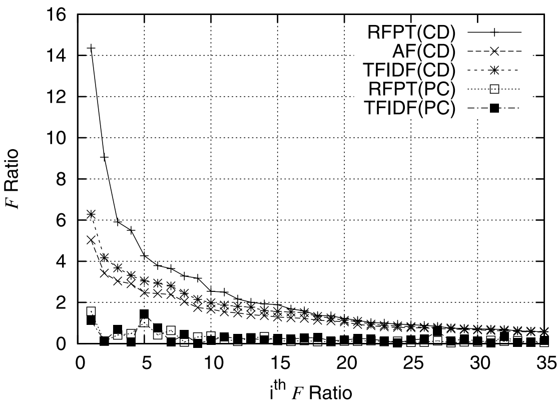


Fig. 2 *F* ratio (a.k.a Fisher's ratio or simply variance ratio) corresponding to eigenvalues. Absolute term frequency (AF), relative frequency with power transformation (RFPT), term frequency weighted by inverse document frequency (TFIDF). The higher the variance ratio the higher the separability. The experiments for PCs were done at 1000 dimensionality.

reasons of comparison, we also give results from term frequency weighted by inverse document frequency (TFIDF) as defined in [10]. The TFIDF results are presented without the transformations as is usually the case in the literature.

Unless otherwise stated, CDA is applied to 1000 principal components (PCs) of RFPT or TFIDF instead of to the direct application to the original features. The reason for choosing 1000 PCs is given in Sect. 3.3. It is also worth mentioning that results for 90 categories are only indicated in Table 1.

4.1 The Impact of Integrated Feature Analysis

Figure 2 shows the *F* ratios of principal components (PCs) and canonical discriminants (CDs) for the between-class documents. This figure shows that CDs have significantly better separability than PCs. It is shown that RFPT has the highest *F* ratios implying better classification effectiveness than that of TFIDF.

In Fig. 3, we see that IFA considerably improves the classification performance. Note that for the PCs, we report the best classification performance at different dimensions. Another point of interest is that IFA performs better at the lowest dimensionality possible. Although TFIDF responded positively to canonical discriminant analysis, the RFPT gave better classification performance than TFIDF.

It is worth mentioning that the direct application of CDA to the original RFPT of 7474 dimensionality achieved 76.4% (micro-averaged F_1). This is significantly worse than with the application of PCA and CDA as well as their integration. This is because of rank deficiency and singularity problems emanating from sample size skewness in relation to high dimensional space. These problems occur when the sample size is smaller than the size of the feature vectors. Thus, the within-class covariance matrices are singular. These problems typically occur in text classification problems. The corpus such as Reuters-21578 contains a considerable number of classes with very small sample size compared to the feature vector dimensionality.

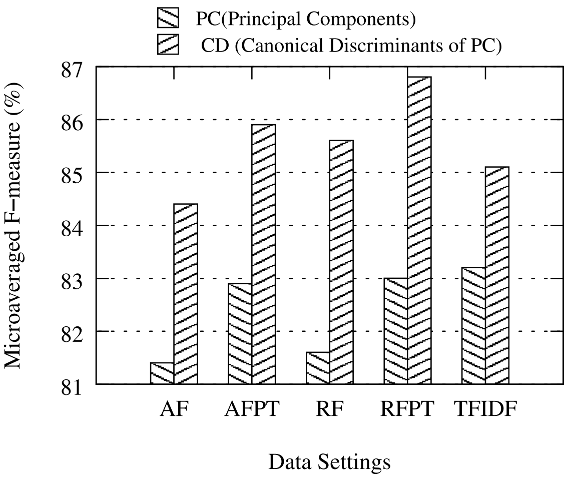


Fig. 3 The effect of IFA. The features include Absolute term frequency (AF) using 500 PCs and 114 CDs; AF after power transformation (AFPT) using 1000 PCs and 114 CDs; Relative term frequency (RF) using 500 PCs and 114 CDs; RF after power transformation (RFPT) using 1000 PCs and 114 CDs; and term frequency weighted by inverse document frequency (TFIDF). TFIDF used 500 PCs and 114 CDs. Note that only the dimensions that gave higher results are considered here.

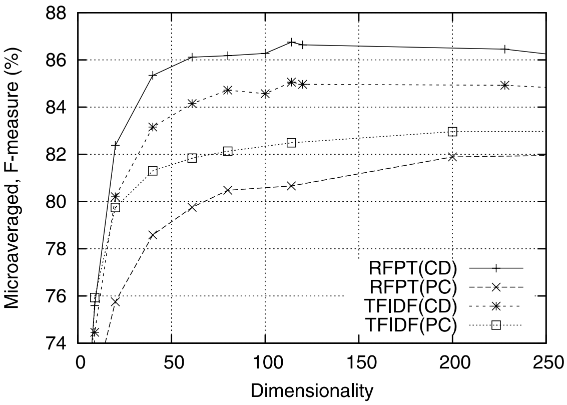
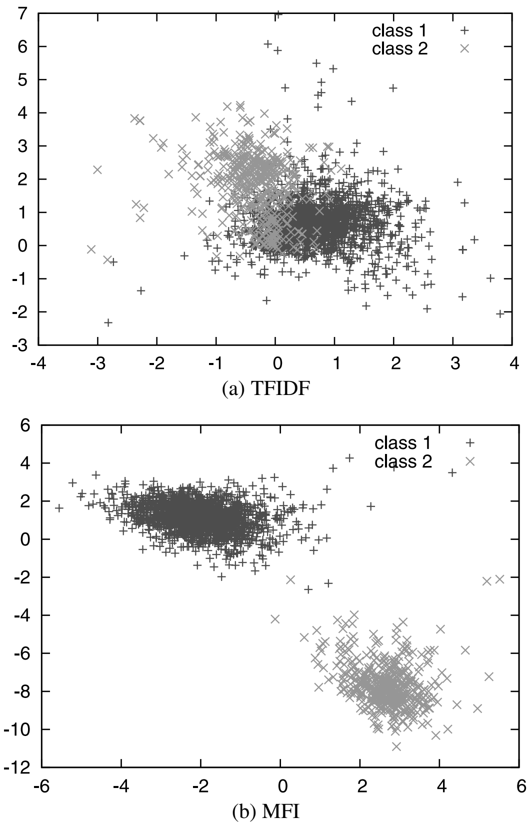


Fig. 4 The impact of IFA in relation to amount of features used. See Fig. 3 for the definition of the abbreviations.

Figure 4 gives the relationship between the dimensionality and the micro-averaged F-measure. The term dimensionality refers to the amount of features used. Comparisons of various data settings have been given. Similarly, it is clear that the IFA considerably improves the classification effectiveness. It can be noted that IFA effect is observably better especially at 114 dimensionality because of the effect of canonical discriminant analysis. It is noted that even when more features were added afterward there was no performance improvement. From Sect. 2.1, it can be recalled that CDA algorithm produces $C - 1$ nonzero generalized eigenvalues leading to equivalent dimensional space that gives the highest performance. This can be the explanation for performance after 114 dimensionality since we used 115 categories in the experiments as described in Sect. 3.3.



(a) TFIDF

(b) MFI

Fig. 5 Class separability for (a) TFIDF and (b) MFI. This effect is illustrated from real data used in experiments i.e., Acquisition category (class 1) and Money-fx category (class 2).

4.2 The Impact of Multiple Feature Integration (MFI)

First of all we show the impact of MFI by giving an example in Fig. 5. We observe that the classes in Fig. 5 (a) are difficult to discriminate. While in Fig. 5 (b) where MFI has been applied, the separability is clear such that the decision boundary can be easily determined. Hence, higher classification performance can be achieved. This indicates that MFI improves the separability between class documents.

We also did multiple feature integration (MFI) of the features called TFIDF and RFPT to generate composite features CF1 and CF2 in Fig. 6. The composite features are defined by

$$CF1 : CD_{144}(PC_{1000}(TFIDF)) \oplus CD_{114}(PC_{1000}(RFPT)), \quad (14)$$

and

$$CF2 : CD_{114}(PC_{1000}(TFIDF) \oplus PC_{1000}(RFPT)). \quad (15)$$

Where $CD_m(X)$ and $PC_m(X)$ denote m canonical discriminants and m principal components of feature X respectively. The symbol $\oplus$ represents the set union (see Sect. 2.5).

We compare the results from TFIDF, RFPT, and those from CF1 and CF2. In Fig. 6, it can be seen that MFI significantly improves the classification performance.

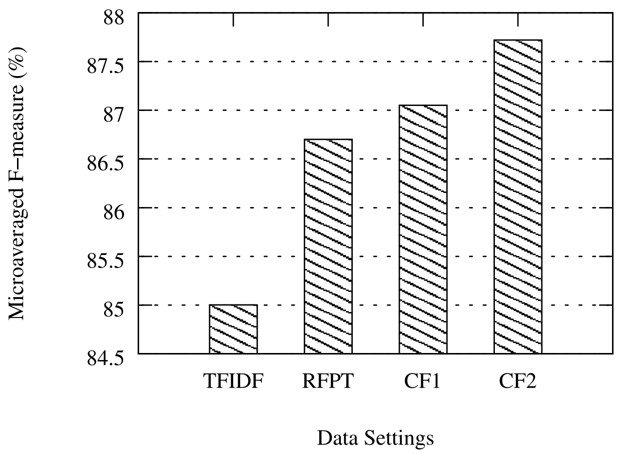


Fig. 6 The impact of multiple feature integration. CF1 = composite feature by the set union of CDs of TFIDF and RFPT (classification at 114 ∗ 2 dimensionality). CF2 = composite feature by the set union of PCs of TFIDF and RFPT followed by CDA (classification at 114 dimensionality).

More importantly, the composite feature CF2 gives the highest performance (87.72% micro-averaged F_1 score at 114 dimensionality). This could be one of the highest score reported in the literature as far as *k*NN classification method and Reuters-21578's ModApte split is concerned.

Table 1 summarizes some results found in the literature in comparison with results presented in this paper. It is also worth noting that most of the works in the literature and those in Table 1 represent textual documents using term weighted vectors commonly called TFIDF. Detailed definitions of the terminologies are given in the indicated references. It is observed that our framework is very effective. This is even obvious when the comparison is done with *k*NN as applied in our experiments. Therefore, it can be said that IFA improves classification performances.

4.3 Classifier's Efficiency Improvements

Nonparametric classifiers such as *k*NN may be slower if subjected to high dimensionality, which is always the case in automatic text classification. Nevertheless, the integration presented in this paper improved the classifier's efficiency with improved classification performance.

Table 2 summarizes the time used for classification at different dimensionalities. The linear transformation column represents the time used to perform linear transformation defined by Eq. (7). It can be seen that the *k*NN time was reduced from 708.3 to 31.6 milliseconds per text, which is about 22 times faster than when using the all words. Similarly, the total time was reduced from 708.3 to 46.3 milliseconds per text, which is about 15 times faster. The encouraging thing is that this efficiency improvement goes along with better classification performance.

In this paper along with other things, we have demonstrated that in contrast to the conventional TFIDF as a way

Table 1 Results (%) in the literature using Reuters-21578's ModApte Split. BEP=Break even point, MI = Mutual Information.

Reference	Data	Classes	Method Summary	Micro-F_1/BEP
This paper	ModApte	115	MFI (CF2), *k*NN, 114 features	87.72
		90	MFI (CF2), *k*NN, 89 features	87.17
Kim et al [6]	ModApte	90	TFIDF, *k*NN, 90 features	86.19
Soucy and Mineau [10]	ModApte	90	Conf.Wt., *k*NN, all features.	86.4
Lam and Han [7]	ModApte	90	TFIDF, GIS, ?features	84.5
Zhang and Oles [9]	ModApte	118	TFIDF;Mod Least Square;10,000features	87.2
			TFIDF;SVM;10,000features	86.5
Yang, Y. [11]	ModApte	90	TFIDF, *k*NN, 24,240features	85
Joachims, T. [12]	ModApte	90	TFIDF, info. Gain,SVM-rbf; 9,962features	86.4
Dumais et al. [14]	ModApte	118	TFIDF,MI,SVM-linear;300features	87.0
Li and Yamanishi [15]	ModApte	90	Binary Words;SVM-linear; thousands?features	84.1
Yang and Liu [16]	ModApte	90	TFIDF;SVM;thousands?features	85.99

Table 2 Classifier's Efficiency (Time in milliseconds per text), LT = linear transformation.

	Dimension	LT Time	*k*NN Time	Total Time
All Features	7474	-	708.3	708.3
PCA	1000	92.4	110.0	202.4
IFA	114	14.7	31.6	46.3

of generating feature vectors, RFPT in an integrated combination approach can be used and its performance exceeds that of its counterparts.

5. Related Works

As far as automatic text classification (ATC) is concerned, and to the best of our knowledge, the proposed integrations are not seen in the literature. Most ATC works consider various techniques in isolation.

All in all, it is worth mentioning that there some works that used the length normalization and power law concepts [8]. However, we note that the approach of the concepts in [8] and ours are not identical. First, they used weighted vectors while we used absolute word frequency and its transformation. Second the length normalization they used is common in the literature and it is different from ours.

Furthermore, Rennie et al. [8] used the concept of power law distribution for the weighted vectors - based on choosing a value of a parameter d, which is added to the weighted vectors, then the result is transformed by computing its *log* value. In contrast, we use the power transformation on absolute word frequency and the relative word frequency.

The weighted vectors such as TFIDF has been conventionally borrowed (e.g. [1], [8]) from information retrieval (IR) and be applied to text classification problems. However, it is unclear whether it is the best choice to represent texts for machine learning systems. A theoretical analysis in [17] has shown that TFIDF is technically best for IR problems.

On the other hand, our approach simultaneously takes care of the problems of dependency on length and sample distribution (see Sects. 2 and 2.6 for the advantages). Furthermore, the problem of high dimensionality is solved in an integrated approach.

Recently, principal component analysis (PCA) has been applied to automatic text classification. This is contrary to the conventional dimension reduction method called latent semantic indexing (LSI) [1], [4]. Lam and Lee [2] did experiments using PCA and neural networks. Among the compared feature reduction techniques, PCA was the best. In ATC, PCA has not been extensively studied in relation to transformed features and their integration.

PCA employs total scatter matrix, which can result in retention of non-discriminative information. Hence, it might be desirable to perform canonical discriminant analysis (CDA) [18] - a common statistical method in other fields of research, but not common in ATC. However, the direct application of CDA may lead to a singularity problem of the within-class covariance matrix. This results from the higher dimensional space than its sample size. Therefore, we use the dominant principal components to which CDA is applied. We also note that the approach we present in this paper is hardly seen in the ATC literature.

In contrary to our approach, Kim et al [6] studied different algorithms for dimension reduction, which is called Linear Discriminant Analysis/Generalized Singular Value Decomposition (LDA/GSVD) algorithm. However, they could not report some of their experimental results. Because their algorithm on Reuters-21578 data collection ran out of memory while computing the GSVD. The reason is that LDA/GSVD is computationally expensive (see page 49 of their paper).

Unlike the previous works, we use transformed features. Particularly, we improve the classification effectiveness by the use of the relative term frequency with the power transformation (RFPT). In addition, we integrate multiple features using the set union concept, which further improves the classification performance.

6. Summary and Future Work

This paper proposes an approach called integrated feature analysis (IFA) which improves text classification performance. Instead of considering feature transformations, principal component and discriminant analysis in isolation, IFA integrates them to achieve higher classification performance. It also proposes multiple feature integration (MFI), which takes into consideration features giving better classification performance from the individual method than when integrating them together.

In short, it can be summarized that we have been able to demonstrate the following:- (1) Integrated feature analysis (IFA) improves the classifiers' performance significantly. (2) Integrating the features by MFI gives superior classification performance without the use of a lot of features. (3) The proposed approach reduces the dimensionality drastically. For instance in our experiments, features were reduced from 7474 to 114 with which we observed highest classification performance. (4) The best classification performance (Micro-averaged $F_1 = 87.72\%$) was achieved by canonical discriminants of RFPT integrated with TFIDF.

We have shown that IFA can improve automatic text classification (ATC). The implication of these results is that, IFA can take a great role in getting higher performance without unnecessarily employing sophisticated classification techniques even at lower dimensionality.

The future direction of this work includes extensive experimental evaluation using more textual samples of more categories including other collections and other classifiers. Collections such as OHSUMED, TREC could be used for further experiments. Other classifiers such SVMs remain for future study.

References

[1] F. Sebastiani, "Machine learning in automated text categorization," ACM Comput. Surv., vol.34, no.1, pp.1–47, 2002.

[2] S. Lam and L. Lee, "Feature reduction for neural network based text categorization," Proc. DASFAA-99, 6th IEEE International Conference on Database Advanced systems for advanced applications, pp.195–202, Hsinchu, TW, 1999.

[3] R. Duda, P. Hart, and D. Stork, Pattern Classification, second ed., John Wiley & Sons, 2001.

[4] M. Wang and J. Nie, "A latent semantic structure model for text classification," Proc. ACM SIGIR workshop on Mathematical/formal methods in information retrieval, Toronto, Canada, 2003.

[5] K. Fukunaga, Introduction to Statistical Pattern Recognition, second ed., Academic Press, Inc, 1990.

[6] H. Kim, P. Howland, and H. Park, "Dimension reduction in text classification with support vector machines," J. Machine Learning Research, vol.6, pp.37–53, 2005.

[7] W. Lam and Y. Han, "Automatic textual document categorization based on generalized instance sets and a metamodel," IEEE Trans. Pattern Anal. Mach. Intell., vol.25, no.5, pp.628–633, 2003.

[8] J. Rennie, L. Shih, J. Teevan, and D. Karger, "Tackling the poor assumptions of naive bayes text classifiers," Proc. 12th International Conference on Machine Learning (ICML), pp.616–623, Washington DC, 2003.

[9] T. Zhang and F. Oles, "Text categorization based on regularized linear classification methods," Information Retrieval Journal, vol.4, pp.5–31, 2001.

[10] P. Soucy and G. Mineau, "Beyond TFIDF weighting for text categorization in the vector space model," Proc. International Joint Conference on Artificial Intelligence (IJCAI), pp.1130–1135, 2005.

[11] Y. Yang, "A study on thresholding strategies for text categorization," Proc. 24th ACM/SIGIR International Conference on Research and Development in Information Retrieval (SIGIR), pp.137–145, 2001.

[12] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," Proc. 10th European Conference on Machine Learning, pp.137–142, 1998.

[13] Y. Yang and X. Liu, "A re-examination of text categorization methods," Proc. Twenty-First International ACM SIGIR Conference on Research and Development in Information Retrieval, pp.42–49, 1999.

[14] S.T. Dumais, J.C. Platt, D. Hecherman, and M. Sahami, "Inductive learning algorithms and representations for text categorization," Proc. 1998 ACM CIKM International Conference on Information and Knowledge Management, Bethesda, Maryland, USA, Nov. 1998, ed. G. Gardarin, J.C. French, N. Pissinou, K. Makki, and L. Bouganim, pp.148–155, ACM, 1998.

[15] H. Li and K. Yamanishi, "Text classification using ESC-based stochastic decision lists," Proc. CIKM-99, 8th ACM International Conference on Information and Knowledge Management, Kansas City, US, pp.122–130, ACM Press, New York, US, 1999.

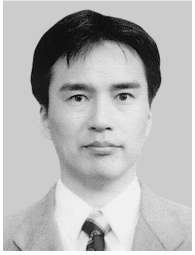
[16] Y. Yang and X. Liu, "A re-examination of text categorization methods," 22nd Annual International SIGIR, Berkley, pp.42–49, Aug. 1999.

[17] K.S. Jones, "A statistical interpretation of term specificity and its application in retrieval," J. Documentation, vol.28, pp.11–12, 1972.

[18] P. Verboon and I.A. van der Lans Pychometrika Journal, vol.59, no.4, pp.48–507, 1994.

Lazaro S.P. Busagala received his Master degree in Information Engineering from Mie University in 2005. He is currently a Ph.D. student at the department of Information Engineering, at the same University. His research interests include text classification, character recognition, document analysis, face and object recognition.

Wataru Ohyama received his Master degree in Engineering from Mie University in 2000. In 2007, he received his PhD in Engineering from Mie University. Since 2000 he has been working as Research Associate for Human Interface Laboratory of the department of Information Engineering of Mie University, Japan. He is currently an Assistant Professor in the same department. He is interested in Research topics such as Medical Image Processing and Medical Signal Processing. He is member of IEEE and IEE Japan.



Tetsushi Wakabayashi received his Master degree from Mie University in 1987, and Ph.D. degree from Nagoya University in 1997. Since 1991, he has been working with Mie University, where he is presently an Associate Professor in the Department of Information Engineering. During 1998-1999, he was a Visiting Scholar at Rensselaer Polytechnic Institute. His research interests are in the area of character recognition, image processing, computer graphics, human computer interaction. He is a member of IPS Japan.

Fumitaka Kimura received his Masters and Ph.D. degrees from Nagoya University in 1975 and 1981, respectively. From 1978 to 1982 he was a faculty member in the Electrical Engineering Department of Nagoya University. Since 1982, he has been working with Mie University, where he is presently a Professor in the Department of Information Engineering. During 1989–1991, he was a Visiting Associate Professor at the University of Michigan-Dearborn in U.S.A. His research interests include in the areas of character recognition, image processing, and document image analysis. He is a member of IPS Japan, MES Japan, and JSAI.