RESEARCH ARTICLE
doi: 10.2306/scienceasia1513-1874.2016.42.052

ScienceAsia 42 (2016): **52–60**

Text classification using similarity measures on intuitionistic fuzzy sets

Peerasak Intarapaiboon

Department of Mathematics and Statistics, Faculty of Science and Technology, Thammasat University, Pathum Thani 12121 Thailand

e-mail: peerasak@mathstat.sci.tu.ac.th

Received 14 Jul 2014
Accepted 16 Aug 2015

ABSTRACT: An intuitionistic fuzzy set (IFS) is an extended version of a fuzzy set and is capable of representing hesitancy degrees. A framework for text classification is presented. Two main challenges are addressed: how to represent documents in terms of IFSs and how to obtain a pattern of each class from such an IFS-based representation. By using some existing similarity measures for IFSs, the proposed framework is applied to two benchmark datasets for text classification. The proposed framework yields satisfactory results when compared to decision tree, *k*-NN, naïve Bayes, and support vector machine classifiers.

KEYWORDS: uncertainty representation, text mining, pattern learning

INTRODUCTION

After the theory of fuzzy sets was proposed by Zadeh[1], many studies have indicated that the theory facilitates solving various real-world problems, especially when dealing with vague information. In fuzzy set theory, membership and nonmembership degrees are complementary, i.e., the sum of both degrees of an element belonging in a fuzzy set is 1. However, there are some situations that the two degrees are not complementary, mainly because of hesitation. Intuitionistic fuzzy set (IFS) was introduced by Atanassov[2] to handle such situations. For representing an IFS, each element is assigned by membership and nonmembership degrees, where the sum of the two degrees does not exceed 1. An IFS is therefore more meaningful than a fuzzy set.

Dengfeng and Chuntian[3] gave the axiomatic definition of similarity measures between IFSs and proposed similarity measures based on high membership and low membership functions. They also paved the way for applying IFS similarity measures to pattern recognition. Liang and Shi[4] showed some counter-intuitive cases obtained from the measures proposed in Ref. 3 and then presented several similarity measures to overcome those cases. Mitchell[5] claimed that the rationale behind unreasonable cases was the weakness of the definition for similarity measures. Thus a stronger definition for grading similarity degree between IFSs was defined. Hung and Yang[6] adopted the Hausdorff distance for developing several similarity measures. Xu[7] introduced the concepts of positive and negative ideal IFSs and extended some similarity measures by assigning weights. The proposed measures were applied to solve multi-attribute decision making problems. Khatibi and Montazer[8] conducted experiments for bacterial classification using a Euclidean-based measure on fuzzy sets, a Euclidean-based measure on IFSs, and a Hausdorff-based measure on IFSs. The results indicated that the both measures on IFSs outperformed others on fuzzy sets. Most similarity measures in the literature are derived from distance measures. As an alternative way, cosine and weighted cosine similarity measures[9] for IFSs were first proposed and applied to a small medical diagnosis problem; after that these measures were modified to satisfy the similarity definition by Hwang and Yang[10]. Reviews of similarity measures for IFSs are presented in Refs. 11, 12.

Text categorization, which involves assigning a textual document to a predefined set of categories, is attracting more attention from researchers. Since this task can be seen as a classification problem from a machine learning point of view, several frameworks using a variety of classification techniques have been proposed. Most classification techniques aim to make a pattern for each category. A new document is then assigned to the category such that its pattern is the most similar to the document. Reviews of text categorization can be found in, e.g., Refs. 13–15. There is little on applying IFSs to

Table 1 Some similarity measures between IFSs.

Author	Expression
Dengfeng[3]	$S_d^p(A,B) = 1-(1/\sqrt[p]{h})\sqrt[p]{\sum_{i=1}^{h}\lvert\varphi_A(i)-\varphi_B(i)\rvert^p}$ where $\varphi_k(i) = (\mu_k(x_i)+1-\nu_k(x_i))/2$, $k=\{A,B\}$, and $1 \leqslant p \leqslant \infty$
Liang[4]	$S_e^p(A,B) = 1-(1/\sqrt[p]{h})\sqrt[p]{\sum_{i=1}^{h}(\frac{1}{2}(\lvert\mu_A(x_i)-\mu_B(x_i)\rvert+\lvert\nu_A(x_i)-\nu_B(x_i)\rvert))^p}$
Mitchell[5]	$S_m^p(A,B) = \frac{1}{2}(\rho_\mu(A,B)+\rho_f(A,B))$ where $\rho_\mu(A,B) = S_d^p(\mu_A(x_i),\mu_B(x_i))$ and $\rho_f(A,B) = S_d^p(1-\nu_A(x_i),1-\nu_B(x_i))$
Xu[7]	$S_Z(A,B) = 1-[(1/2h)\sum_{i=1}^{h}((\mu_A(x_i)-\mu_B(x_i))^p+(\nu_A(x_i)-\nu_B(x_i))^p+(\pi_A(x_i)-\pi_B(x_i))^p)]^{1/p}$
Julian[17]	$S_n^p(A,B) = 1-\sqrt[p]{\sum_{i=1}^{h}w_i(\lvert\mu_A(x_i)-\mu_B(x_i)\rvert)^p}-\sqrt[p]{\sum_{i=1}^{h}w_i(\lvert\nu_A(x_i)-\nu_B(x_i)\rvert)^p}$ with $w_i \geqslant 0$ and $\sum_{i=1}^{h}w_i = 1$ and $p \geqslant 1$
Ye[9]	$S_C(A,B) = (1/h)\sum_{i=1}^{h}(\mu_A(x_i)\mu_B(x_i)+\nu_A(x_i)\nu_B(x_i))/(\sqrt{\mu_A^2(x_i)+\nu_A^2(x_i)}\sqrt{\mu_B^2(x_i)+\nu_B^2(x_i)})$
Hwang and Yang[10]	$S_S(A,B) = \frac{1}{3}(S_C(A,B)+C_{\mathrm{IFS}}^*(A,B)+C_{\mathrm{IFS}}^{**}(A,B))$ where $C_{\mathrm{IFS}}^*(A,B) = \frac{1}{h}\sum_{i=1}^{h}\frac{\varphi_A(x_i)\varphi_B(x_i)+\nu_A(x_i)\nu_B(x_i)}{\sqrt{\varphi_A^2(x_i)+\nu_A^2(x_i)}\sqrt{\varphi_B^2(x_i)+\nu_B^2(x_i)}}$, with $\varphi_k(x_i) = \frac{1}{2}(1+\nu_k(x_i)-\nu_k(x_i))$, $k=A$, B and $C_{\mathrm{IFS}}^{**}(A,B) = \frac{1}{h}\sum_{i=1}^{h}\frac{(1-\mu_A(x_i))(1-\mu_B(x_i))+(1-\nu_A(x_i))(1-\nu_B(x_i))}{\sqrt{(1-\mu_A(x_i))^2+(1-\nu_A(x_i))^2}\sqrt{(1-\mu_B(x_i))^2+(1-\nu_B(x_i))^2}}$

text categorization in the literature. Szmidt and Kacprzyk[16] proposed a strategy for feature selection in text categorization using the concept of IFSs. No framework obviously shows the benefit of IFS similarity measures to text categorization.

To shed light on this study direction, a framework for text classification based on IFS similarity measures is presented. We address how to represent a document in terms of an IFS and how to obtain a pattern of each class from such an IFS-based representation. Our framework is then evaluated on two benchmark datasets of text classification and compared to other traditional text classification methods.

INTUITIONISTIC FUZZY SETS AND THEIR SIMILARITY MEASURES

In this section, we present some basic concepts for IFSs and their similarity measures. The following notation is used hereinafter: $X = \{x_1, x_2, \ldots, x_h\}$ is a discrete universe of discourse; IFS(X) is the collection of all IFSs of X. An intuitionistic fuzzy set A in IFS(X) is defined as follows:

$$A = \{\langle x_i, \mu_A(x_i), \nu_A(x_i)\rangle \mid x_i \in X\}$$

which is characterized by a membership function $\mu_A(x_i)$ and a nonmembership function $\nu_A(x_i)$. The two functions are defined as follows:

$$\mu_A : X \to [0,1],$$

$$\nu_A : X \to [0,1],$$

such that

$$0 \leqslant \mu_A(x_i)+\nu_A(x_i) \leqslant 1, \quad \forall x_i \in X.$$

Another degree, $\pi_A(x_i)$, the hesitancy degree of x_i belonging to A, is defined as

$$\pi_A(x_i) = 1-\mu_A(x_i)-\nu_A(x_i). \tag{1}$$

Definition 1 Let S be a real-valued function such that S : IFS(X) × IFS(X) → [0,1]. S is called a similarity measure if, for all A,B,C in IFS(X), it satisfies the following conditions: $S(A,B) = S(B,A)$; $S(A,B) = 1$ iff $A = B$; if $A \subseteq B \subseteq C$, then $S(A,C) \leqslant S(A,B)$ and $S(A,C) \leqslant S(B,C)$.

Assuming $A = \{\langle x_i, \mu_A(x_i), \nu_A(x_i)\rangle \mid x_i \in X\}$ and $B = \{\langle x_i, \mu_B(x_i), \nu_B(x_i)\rangle \mid x_i \in X\}$ are IFSs, Table 1 highlights some similarity measures between IFSs. S_d^p, S_e^p, S_m^p, S_Z, and S_n^p are distance-based measures, since each of them is defined as one minus the distance. S_C and S_S are cosine-based measures.

AN IFS-BASED FRAMEWORK FOR TEXT CLASSIFICATION

The proposed IFS-based framework for text classification is outlined in Fig. 1. The details of this framework are discussed below.

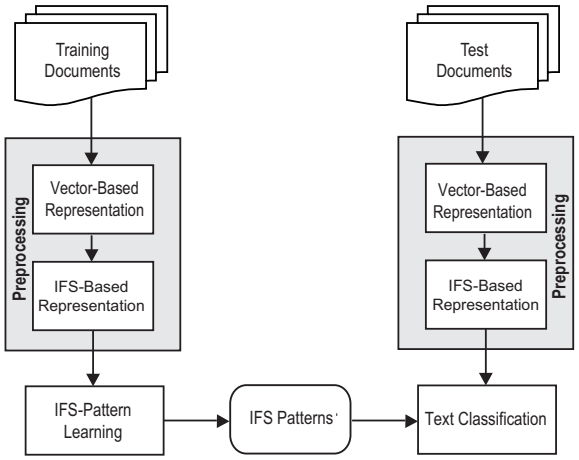

Fig. 1 An overview of the proposed framework.

Preprocessing

In the preprocessing step, a document is represented as a bag of words, one of the basic methods for representing a document. The bag of words is used to form a vector for representing a document using the frequency of each word determined as a relevant feature for categorization. Such words can be selected by some feature selection techniques, e.g., information gain and gain ratio. Assume h words, i.e., $w_1, w_2, \ldots, w_h$, are selected for this representation. Document d_i can be represented by

$$V_i = (n_{i,1}, n_{i,2}, \ldots, n_{i,h}), \tag{2}$$

where $n_{i,j}$ denotes the number of occurrences of w_j in d_i, $j = 1, 2, \ldots, h$.

To convert a bag of words vector to an IFS, we propose one method of which the conceptual idea is explained as follows: Suppose $A_i = \{\langle x_1, \mu_i(x_1), \nu_i(x_1)\rangle \ldots \langle x_h, \mu_i(x_h), \nu_i(x_h)\rangle\}$ is an IFS for the vector V_i in (2). In this work, $\mu_i(x_j)$ represents a confidence level to say that word w_j occurs in document d_i, while $\nu_i(x_j)$ is a confidence level to say that word w_j does not occur in document d_i. The next example gives more details.

Example 1 Assume there are three documents, i.e., d_1, d_2, and d_3, four feature words, i.e., w_1, w_2, w_3, and w_4, and the bag of words-based vectors for these documents, respectively, are

$$V_1 = (10, 2, 5, 0), V_2 = (5, 9, 5, 1), V_3 = (1, 8, 4, 3).$$

Since $n_{1,1} > n_{2,1} > n_{3,1}$, the confidence level to say that w_1 occurs in d_1 should be more than for d_2 and d_3. It implies that $\mu_1(x_1) > \mu_2(x_1) > \mu_3(x_1)$ and $\nu_1(x_1) < \nu_2(x_1) < \nu_3(x_1)$. Consider w_2. We should have $\mu_1(x_1) < \mu_2(x_1) \approx \mu_3(x_1)$ and $\nu_1(x_1) > \nu_2(x_1) \approx \nu_3(x_1)$. In the case of w_3 where $n_{1,3} = n_{2,3} \approx n_{3,3}$, we have low confidence to assign high values for the membership and nonmembership degrees of every document. Thus we set $\mu_1(x_3) = \mu_2(x_3) = 0.2$, $\mu_3(x_3) = 0.15$, $\nu_1(x_3) = \nu_2(x_3) = 0.2$, and $\nu_3(x_3) = 0.15$.

Based on the ideas discussed above, the process of transformation will be formally explained. Given the universe of discourse $X = \{\mathrm{HF}_1, \mathrm{HF}_2, \ldots, \mathrm{HF}_h\}$. ($\mathrm{HF}_i$ has an internal meaning as a high frequency of w_i.) Every value $n_{i,j}$ in the vector-based representation of the document i is then expressed in terms of the three degrees of HF_j as in the following steps:

(i) $n_{i,j}$ is normalized by

$$z_{i,j} = \frac{n_{i,j} - \bar{X}_j}{s_j}, \tag{3}$$

where $\bar{X}_j$ and s_j are the mean and the standard deviation, respectively, of the feature word w_j.

(ii) Denoted by $\mu_{i,j}$, a membership degree of the document i with respect to HF_j is determined by a weighted sigmoid function:

$$\mu_{i,j} = \frac{r_j}{1 + \mathrm{e}^{-z_{i,j}}}, \tag{4}$$

where r_j is a weight for HF_j.

(iii) Denoted by $\nu_{i,j}$, the nonmembership degree of the document i with respect to HF_j is calculated by a weighted sigmoid function:

$$\nu_{i,j} = \frac{r_j^*}{1 + \mathrm{e}^{z_{i,j}}}, \tag{5}$$

where r_j^* is a weight for HF_j.

(iv) Denoted by $\pi_{i,j}$, the hesitancy degree of the document i with respect to HF_j is calculated by (1), i.e.,

$$\pi_{i,j} = 1 - \mu_{i,j} - \nu_{i,j}.$$

Pattern learning

This section presents a procedure for learning patterns of predefined classes in terms of IFSs. Assume that there are l classes referred to as $C_1, C_2, \ldots, C_l$; m training documents denoted by $d_1, d_2, \ldots, d_m$; and h word features, namely, $w_1, w_2, \ldots, w_h$. A pattern for class C_k, denoted by P_k, is defined by

$$P_k = \{(\mathrm{HF}_j, \bar{\mu}_{kj}, \bar{\nu}_{kj})\}_{j=1}^{h},$$

where $\bar{\mu}_{kj}$ and $\bar{\nu}_{kj}$ are the average values of membership and nonmembership, respectively, of the

Table 2 A training set used in Example 2.

Document	w_1	w_2	w_3	Class
d_1	15	2	7	C_1
d_2	10	3	5	C_1
d_3	2	9	6	C_2
d_4	3	11	5	C_2
d_5	3	9	4	C_2

word feature w_j observed from d_i belonging to C_k. More precisely,

$$\bar{\mu}_{kj} = \frac{\sum_{i=1}^{m}(\mu_{ij}\chi_k(d_i))}{\sum_{i=1}^{m}\chi_k(d_i)}, \ \bar{\nu}_{kj} = \frac{\sum_{i=1}^{m}(\nu_{ij}\chi_k(d_i))}{\sum_{i=1}^{m}\chi_k(d_i)},$$

where

$$\chi_k(d_i) = \begin{cases} 1, & d_i \in C_k, \\ 0, & \text{otherwise.} \end{cases}$$

Text classification

To assign a proper class to a test document d_t, we represent d_t in terms of an IFS by the same values of parameters during the training process. Intuitively, d_t is grouped into class C' such that its pattern is closest to the IFS representation of d_t. More formally,

$$C' = \arg\max_{C_k}\{\mathrm{Sim}(P_k, \mathrm{IFS}_{d_t})\},$$

where Sim is a similarity measure between IFSs; P_k is an IFS-based pattern of class C_k; and IFS_{d_t} is the IFS representation of d_t.

Example 2 This example shows the details of the proposed framework. Assume that there are five training documents, i.e., d_1, d_2, d_3,d_4, and d_5; three feature words, i.e., w_1, w_2, and w_3; and two document classes, referred to as C_1 and C_2. The frequency of each word and the class of each document are depicted in Table 2. For example, d_1 has 15, 2, and 7 occurrences of w_1, w_2, and w_3, respectively, and its class is C_1.

To represent those documents as IFSs, we start by defining the universe $X = \{\mathrm{HF}_1, \mathrm{HF}_2, \mathrm{HF}_3\}$. From Table 2, $\bar{X}_1$, $\bar{X}_2$, $\bar{X}_3$, s_1, s_2, and s_3 are 6.6, 6.8, 5.4, 5.68, 4.02, and 1.14, respectively. By (3), (4), and (5) with $r_j = r_j^* = 0.9$, we obtain IFSs corresponding to these documents as

$$\begin{aligned}
\mathrm{IFS}_1 &= \{\langle \mathrm{HF}_1, 0.73, 0.17\rangle, \langle \mathrm{HF}_2, 0.21, 0.69\rangle, \\
&\qquad \langle \mathrm{HF}_3, 0.72, 0.18\rangle\}, \\
\mathrm{IFS}_2 &= \{\langle \mathrm{HF}_1, 0.58, 0.32\rangle, \langle \mathrm{HF}_2, 0.25, 0.65\rangle, \\
&\qquad \langle \mathrm{HF}_3, 0.37, 0.53\rangle\}, \\
\mathrm{IFS}_3 &= \{\langle \mathrm{HF}_1, 0.28, 0.62\rangle, \langle \mathrm{HF}_2, 0.57, 0.33\rangle, \\
&\qquad \langle \mathrm{HF}_3, 0.57, 0.33\rangle\}, \\
\mathrm{IFS}_4 &= \{\langle \mathrm{HF}_1, 0.31, 0.59\rangle, \langle \mathrm{HF}_2, 0.67, 0.23\rangle, \\
&\qquad \langle \mathrm{HF}_3, 0.37, 0.53\rangle\}, \\
\mathrm{IFS}_5 &= \{\langle \mathrm{HF}_1, 0.31, 0.59\rangle, \langle \mathrm{HF}_2, 0.57, 0.33\rangle, \\
&\qquad \langle \mathrm{HF}_3, 0.20, 0.70\rangle\}.
\end{aligned}$$

To construct class patterns, the documents are grouped by their classes and then the class pattern can be obtained by averaging membership and non-membership levels. The patterns are

$$\begin{aligned}
P_1 &= \{\langle \mathrm{HF}_1, 0.66, 0.24\rangle, \langle \mathrm{HF}_2, 0.23, 0.67\rangle, \\
&\qquad \langle \mathrm{HF}_3, 0.55, 0.35\rangle\}, \\
P_2 &= \{\langle \mathrm{HF}_1, 0.30, 0.60\rangle, \langle \mathrm{HF}_2, 0.60, 0.30\rangle, \\
&\qquad \langle \mathrm{HF}_3, 0.38, 0.52\rangle\}.
\end{aligned}$$

Suppose that d_t is a new document to be classified in which the frequencies of feature words w_1, w_2, and w_3 are 10, 5, and 5, respectively. Using the same values of $\bar{X}_j$ and s_j in the training process, d_t can be represented as an IFS:

$$\begin{aligned}
\mathrm{IFS}_{d_t} = \{&\langle \mathrm{HF}_1, 0.58, 0.32\rangle, \\
&\langle \mathrm{HF}_2, 0.39, 0.51\rangle, \langle \mathrm{HF}_3, 0.39, 0.51\rangle\}.
\end{aligned}$$

If S_C in Table 1 is used to calculate the similarity between P_k and IFS_{d_t}, where $k = 1, 2$, then we have

$$S_C(P_1, \mathrm{IFS}_{d_t}) = 0.96, \quad S_C(P_2, \mathrm{IFS}_{d_t}) = 0.91.$$

Hence d_t is classified as C_1.

DATASETS AND EXPERIMENTAL SETTINGS

Datasets

In this paper, two news datasets[18], namely, BBC and BBCSport, were used for our experiments. The BBC dataset is constructed from 2225 news articles in 5 topical areas. With basic preprocessing steps including stemming, stop-word removal, and low term frequency filtering, 9635 words are obtained for representing each article. The BBCSport dataset contains 737 sports news articles on 5 areas. Each article is represented by the frequency of 4613 words obtained using the same processes applied to BBC. Table 3 summarizes important characteristics of each dataset.

Table 3 Data set characteristics.

Dataset	#articles	#words	Class distribution
BBC	2225	9635	474:374:434:506:437
BBCSport	737	4613	101:124:265:147:100

Experimental schema and parameter setting

As seen in Table 3, the feature-space dimensions of both datasets are very high. In order to reduce the dimensions, two measures, the *gain ratio* (GR) and *information gain* (IG) were applied for evaluating feature relevance. The top 20% and 30% of features ranked according to their GR and IG values for BBC and the top 50% and 80% of those for BBCSport were selected.

To retain the independence of the data in use, 5-fold cross-validation (CV) was adopted for evaluating the proposed framework. In our experiment, we explore two strategies for setting the weights r_j and r_j^* in (4) and (5):

(i) A varied grid-based method, referred to as VG, varies the weights from 0.6–1 with increments of 0.1. If 0.8 and 0.9 are set for r_j and r_j^*, respectively, then r_j is fixed at 0.8 while r_j^* is 0.9 for all j. There are thus 25 alternatives which come from the combinations of 0.6–1 of r_j and 0.6–1 of r_j^*.

(ii) An adaptive method, referred to as AD, sets the two weights based on statistical characteristics of feature words by

$$r_j = r_j^* = \left|\frac{1-s_j}{1+s_j}\right|, \tag{6}$$

where s_j is the standard deviation of the number of occurrences of w_j in training documents.

In this paper, all IFS similarity measures listed in Table 1 were used when the parameter p was set to 2 for S_d^p, S_e^p, S_m^p, S_Z, and S_n^p, and each w_i of S_n^p was set equal to the reciprocal of the number of features.

Evaluation metric

Given a test set, l denotes the number of classes, and n_i denotes the number of documents in class i. When an experiment is done, some preliminary performance measures with respect to class i, i.e., precision (P_i), recall (R_i), and F-measure (F_i), are calculated. They are defined by

$$P_i = \frac{\mathrm{TP}_i}{\mathrm{TP}_i + \mathrm{FP}_i},\ R_i = \frac{\mathrm{TP}_i}{n_i},\ F_i = \frac{2R_iP_i}{R_i+P_i},$$

where TP_i (true positive) refers to the number of documents correctly classified as class i; with $i \neq j$, FN_i (false negative) refers to the number of documents in class i classified as class j; and FP_i (false positive) to the number of documents in j classified as i. Since every subdataset for the 5-fold CV contains 5 classes and the numbers of documents belonging in the classes are not equal, all experiments are multi-class imbalance scenarios. To show the average performance over all classes, we use the mean F-measure (MFM) and macro geometric average (MAvG), which are widely used in this situation[19], were applied. These two evaluation metrics are defined as follows:

$$\mathrm{MFM} = \frac{\sum_{i=1}^{l} F_i}{l}, \quad \mathrm{MAvG} = \sqrt[l]{\Pi_{i=1}^{l} R_i}.$$

To present a level of false classification for the proposed framework, AvD, the average difference between the similarity degree with the correct classes and that with the predicted classes over all false predictions, is calculated as

$$\mathrm{AvD} = \sum_{d_t \in \mathrm{FC}} \big(|\mathrm{Sim}(\mathrm{IFS}(d_t), \mathrm{IFS}(\mathrm{Cor}_{d_t})) - \mathrm{Sim}(\mathrm{IFS}(d_t), \mathrm{IFS}(\mathrm{Prt}_{d_t}))|\big) / |\mathrm{FC}|,$$

where d_t is a test document, $\mathrm{IFS}(d_t)$ is the IFS representing d_t, $\mathrm{IFS}(\mathrm{Cor}_{d_t})$ is the IFS representing the correct class of d_t, $\mathrm{IFS}(\mathrm{Prt}_{d_t})$ is the IFS representing the predicted class of d_t, and FC is the set of false classified documents.

EXPERIMENTAL RESULTS AND DISCUSSION

The first experiment aims to compare performances of the varied grid-based and adaptive strategies for weight setting. The second experiment compares our IFS-based framework with traditional classification methods including decision tree, k-NN, naïve Bayes, and support vector machine.

Comparison between the VG and AD weight setting methods

Using the similarity measures in Table 1, Tables 4 and 5 give the evaluation results on BBC when the gain ratio and information gain are used, respectively. Likewise, Tables 6 and 7 give results on BBCSport. In each table, the first column presents the percentage of selected features, the second means the strategy of weight setting, and the others show experimental results in terms of MFM (in percent), MAvG (in percent), and AvD. A value in parentheses after an evaluation score presents the uncertainty of the last figure in terms of the standard deviation

Table 4 Comparison between the VG and AD methods on BBC with top N% of features ranked by GR.

N	Sim	W	MFM(SD)	MAvG(SD)	AvD(SD)
20	S_d^p	AD	93.2(4)	92.5(4)	$3.3\,10^{-4}(5)\uparrow$
		VG	$94.3(9)^{[1,1]}$	$93(1)^{[1,1]}$	$7.4\,10^{-4}(4)^{[0.6,0.6]}$
	S_e^p	AD	93.2(4)	92.6(4)	$3.3\,10^{-4}(5)\uparrow$
		VG	$94.3(9)^{[1,1]}$	$93.9(1)^{[1,1]}$	$7.4\,10^{-4}(4)^{[0.6,0.6]}$
	S_m^p	AD	93.2(4)	92.6(4)	$1.7\,10^{-4}(2)\uparrow$
		VG	$94.3(9)^{[1,1]}$	$93.9(1)^{[1,1]}$	$3.7\,10^{-4}(2)^{[0.6,0.6]}$
	S_z	AD	93.2(4)	92.6(4)	$3.3\,10^{-4}(5)\uparrow$
		VG	$94.3(9)^{[1,1]}$	$93.9(1)^{[1,1]}$	$7.4\,10^{-4}(4)^{[0.6,0.6]}$
	S_n^p	AD	93.2(4)	92.6(4)	$6.6\,10^{-4}(1)\uparrow$
		VG	$94.3(9)^{[1,1]}$	$93.9(1)^{[1,1]}$	$1.5\,10^{-3}(1)^{[0.6,0.6]}$
	S_C	AD	92.8(1)	92.1(2)	$1.7\,10^{-4}(1)$
		VG	$94.0(9)^{[0.6,1]}$	$92.6(2)^{[0.6,1]}$	$1.3\,10^{-4}(7)^{[1,0.6]}$
	S_S	AD	85.1(1) $\downarrow$	82.42(2) $\downarrow$	$3.0\,10^{-5}(4)\uparrow$
		VG	$91.1(2)^{[1,0.6]}$	$89.2(2)^{[1,0.6]}$	$5.0\,10^{-5}(6)^{[1,0.6]}$
30	S_d^p	AD	94.7(6)	94.4(7)	$2.6\,10^{-4}(3)\uparrow$
		VG	$94.7(6)^{[1,1]}$	$94.4(6)^{[1,1]}$	$5.5\,10^{-4}(6)^{[0.6,0.6]}$
	S_e^p	AD	94.7(6)	94.4(7)	$2.6\,10^{-4}(3)\uparrow$
		VG	$94.7(6)^{[1,1]}$	$94.4(6)^{[1,1]}$	$5.5\,10^{-4}(6)^{[0.6,0.6]}$
	S_m^p	AD	94.7(6)	94.4(7)	$1.3\,10^{-4}(1)\uparrow$
		VG	$94.7(6)^{[1,1]}$	$94.4(6)^{[1,1]}$	$2.7\,10^{-4}(3)^{[0.6,0.6]}$
	S_z	AD	94.7(6)	94.4(7)	$2.6\,10^{-4}(3)\uparrow$
		VG	$94.7(6)^{[1,1]}$	$94.4(6)^{[1,1]}$	$5.5\,10^{-4}(6)^{[0.6,0.6]}$
	S_n^p	AD	94.7(6)	94.4(7)	$5.2\,10^{-4}(7)\uparrow$
		VG	$94.7(6)^{[1,1]}$	$94.4(6)^{[1,1]}$	$1.1\,10^{-3}(1)^{[0.6,0.6]}$
	S_C	AD	93.4(1) $\downarrow$	93.0(1) $\downarrow$	$2.2\,10^{-4}(1)$
		VG	$95.2(8)^{[0.6,1]}$	$95.0(2)^{[0.6,1]}$	$1.7\,10^{-4}(1)^{[1,0.6]}$
	S_S	AD	84.8(2) $\downarrow$	82.8(2) $\downarrow$	$6.0\,10^{-5}(7)\uparrow$
		VG	$89.7(1)^{[0.6,1]}$	$87.8(1)^{[0.6,1]}$	$7.0\,10^{-5}(1)^{[1,0.6]}$

Table 5 Comparison between the VG and AD methods on BBC with top N% of features ranked by IG.

N	Sim	W	MFM(SD)	MAvG(SD)	AvD(SD)
20	S_d^p	AD	94.3(9)	94.0(1)	$3.6\,10^{-4}(5)\uparrow$
		VG	$93.9(1)^{[1,1]}$	$93.6(1)^{[1,1]}$	$7.1\,10^{-4}(1)^{[0.6,0.6]}$
	S_e^p	AD	94.4(9)	94.0(1)	$3.6\,10^{-4}(5)\uparrow$
		VG	$93.9(1)^{[1,1]}$	$93.6(1)^{[1,1]}$	$7.1\,10^{-4}(1)^{[0.6,0.6]}$
	S_m^p	AD	94.4(9)	94.0(1)	$1.8\,10^{-4}(3)\uparrow$
		VG	$93.9(1)^{[1,1]}$	$93.6(1)^{[1,1]}$	$3.6\,10^{-4}(6)^{[0.6,0.6]}$
	S_z	AD	94.4(9)	94.0(1)	$3.6\,10^{-4}(5)\uparrow$
		VG	$93.9(1)^{[1,1]}$	$93.6(1)^{[1,1]}$	$7.1\,10^{-4}(1)^{[0.6,0.6]}$
	S_n^p	AD	94.4(9)	94.0(1)	$7.3\,10^{-4}(1)\uparrow$
		VG	$93.9(1)^{[1,1]}$	$93.6(1)^{[1,1]}$	$1.4\,10^{-3}(2)^{[0.6,0.6]}$
	S_C	AD	92.3(2) $\downarrow$	91.8(2) $\downarrow$	$3.7\,10^{-4}(5)$
		VG	$94.6(1)^{[0.6,1]}$	$93.0(1)^{[0.6,1]}$	$3.0\,10^{-4}(4)^{[1,0.6]}$
	S_S	AD	82.7(2) $\downarrow$	80.3(3) $\downarrow$	$1.0\,10^{-4}(1)\uparrow$
		VG	$88.3(1)^{[0.6,1]}$	$85.5(2)^{[0.6,1]}$	$1.6\,10^{-4}(2)^{[1,0.6]}$
30	S_d^p	AD	94.5(9)	94.2(1)	$2.7\,10^{-4}(1)\uparrow$
		VG	$94.0(1)^{[1,1]}$	$93.7(1)^{[1,1]}$	$5.7\,10^{-4}(9)^{[0.6,0.6]}$
	S_e^p	AD	94.5(9)	94.2(1)	$2.7\,10^{-4}(1)\uparrow$
		VG	$94.0(1)^{[1,1]}$	$93.7(1)^{[1,1]}$	$5.7\,10^{-4}(9)^{[0.6,0.6]}$
	S_m^p	AD	94.5(9)	94.2(1)	$1.4\,10^{-4}(9)\uparrow$
		VG	$94.0(1)^{[1,1]}$	$93.7(1)^{[1,1]}$	$2.8\,10^{-4}(4)^{[0.6,0.6]}$
	S_z	AD	94.5(9)	94.2(1)	$2.7\,10^{-4}(1)\uparrow$
		VG	$94.0(1)^{[1,1]}$	$93.7(1)^{[1,1]}$	$5.7\,10^{-4}(9)^{[0.6,0.6]}$
	S_n^p	AD	94.5(9)	94.2(1)	$5.4\,10^{-4}(3)\uparrow$
		VG	$94.0(1)^{[1,1]}$	$93.7(1)^{[1,1]}$	$1.1\,10^{-3}(1)^{[0.6,0.6]}$
	S_C	AD	92.3(2) $\downarrow$	91.8(2) $\downarrow$	$2.5\,10^{-4}(3)\uparrow$
		VG	$94.7(1)^{[0.6,1]}$	$93.1(2)^{[0.6,1]}$	$2.0\,10^{-4}(2)^{[1,0.6]}$
	S_S	AD	83.6(2) $\downarrow$	81.4(3) $\downarrow$	$7.0\,10^{-5}(7)\uparrow$
		VG	$88.7(1)^{[0.6,1]}$	$85.7(2)^{[0.6,1]}$	$9.0\,10^{-5}(1)^{[1,0.6]}$

(SD). For instance, 93.2(4) means 93.2 ± 0.4, while $3.3\,10^{-4}(5)$ means 0.00033 ± 0.00005.

For brevity, only the best value of each metric is reported for the VG method. The superscripts $[x, y]$ denote the values of r_j and r_j^*, respectively, that yield such the highest value.

From the tables, we can see that the experimental results from AD are comparable to those from VG. For the tests on BBC using gain ratio for feature selection (Table 4), regardless of similarity measures, the MFM and MAvG scores resulting from our framework with VG are slightly higher than those from our framework with AD. In contrast, considering the AvD values, the framework with AD outperforms the framework with VG with the average difference of $3.20\,10^{-4}$. With information gain (Table 5), all the results obtained from the distance-based measures, i.e., S_d^p, S_e^p, S_m^p, S_z, and S_n^p with AD are better than those with VG. The cosine-based measures, i.e., S_C and S_S, with AD yield lower values of MFM and MAvG than the same measures with VG. Likewise, on BBCSport (Tables 6 and 7), we observe that the experimental results from AD are comparable to those from VG.

A two-tailed paired t-test at 5% for each metric was performed. In the rows presenting the AD results of the experimental tables, $\uparrow$ or $\downarrow$ indicates that AD is significantly better or worse, respectively, than VG, while no mark means no significant difference. Significant differences occur mostly for assessment with AvD, especially testing BBC. Only two similarity measures, i.e., S_C and S_S, produces significant differences for MFM and MAvG. Even though we cannot conclude which method is more efficient, the optimal results of VG are seen to depend on evaluation metrics and similarity measures.

Comparison with other methods

The proposed framework was also compared with classification by four standard models, i.e., decision tree (DT) using C4.5, naïve Bayes (NB), k-nearest neighbour (k-NN), and support vector machine (SVM) based on the RBF kernel. The Weka machine learning suite (www.cs.waikato.ac.nz/ml/weka) was employed for classifier learning and evaluation, using its default parameters. As observed during the learning process, 3-NN performed slightly better than 1-NN, 5-NN, and 7-NN on our

Table 6 Comparison between the VG and AD methods on BBCSport with top N% of features ranked by GR.

N	Sim	W	MFM(SD)	MAvG(SD)	AvD(SD)
50	S_d^p	AD	93.9(1)	92.5(2)	$1.0\,10^{-3}$(3)
		VG	93.7(6)$^{[1,1]}$	91.7(7)$^{[1,1]}$	$0.8\,10^{-3}$(2)$^{[0.6,0.6]}$
	S_e^p	AD	93.9(1)	92.5(2)	$1.0\,10^{-3}$(3)
		VG	93.7(6)$^{[1,1]}$	91.7(7)$^{[1,1]}$	$0.8\,10^{-3}$(2)$^{[0.6,0.6]}$
	S_m^p	AD	93.9(1)	92.5(2)	$0.5\,10^{-3}$(2)
		VG	93.7(6)$^{[1,1]}$	91.7(7)$^{[1,1]}$	$4.1\,10^{-4}$(8)$^{[0.6,0.6]}$
	S_z	AD	93.9(1)	92.5(2)	$1.0\,10^{-3}$(3)
		VG	93.7(6)$^{[1,1]}$	91.7(7)$^{[1,1]}$	$0.8\,10^{-3}$(2)$^{[0.6,0.6]}$
	S_n^p	AD	93.9(1)	92.5(2)	$2.1\,10^{-3}$(7)
		VG	93.7(6)$^{[1,1]}$	91.7(7)$^{[1,1]}$	$1.6\,10^{-3}$(3)$^{[0.6,0.6]}$
	S_C	AD	91.3(6) $\downarrow$	88.4(1) $\downarrow$	$3.6\,10^{-4}$(3) $\downarrow$
		VG	93.7(8)$^{[0.6,1]}$	89.9(1)$^{[0.6,1]}$	$2.8\,10^{-4}$(3)$^{[1,0.6]}$
	S_S	AD	90.2(1) $\uparrow$	87.2(2) $\uparrow$	$2.6\,10^{-4}$(3) $\downarrow$
		VG	88.0(1)$^{[1,0.6]}$	82.6(2)$^{[1,0.6]}$	$1.2\,10^{-4}$(1)$^{[1,0.6]}$
80	S_d^p	AD	93.7(1)	92.2(2)	$0.7\,10^{-3}$(2)
		VG	94.1(5)$^{[1,1]}$	92.2(8)$^{[1,1]}$	$5.9\,10^{-3}$(3)$^{[0.6,0.6]}$
	S_e^p	AD	93.7(1)	92.2(2)	$0.7\,10^{-3}$(2)
		VG	94.1(5)$^{[1,1]}$	92.2(8)$^{[1,1]}$	$5.9\,10^{-3}$(3)$^{[0.6,0.6]}$
	S_m^p	AD	93.7(1)	92.2(2)	$0.3\,10^{-3}$(1)
		VG	94.1(5)$^{[1,1]}$	92.2(8)$^{[1,1]}$	$2.9\,10^{-4}$(1)$^{[0.6,0.6]}$
	S_z	AD	93.7(1)	92.2(2)	$0.7\,10^{-3}$(2)
		VG	94.1(5)$^{[1,1]}$	92.2(8)$^{[1,1]}$	$5.9\,10^{-4}$(3)$^{[0.6,0.6]}$
	S_n^p	AD	93.7(1)	92.2(2)	$1.5\,10^{-3}$(5)
		VG	94.1(5)$^{[1,1]}$	92.2(8)$^{[1,1]}$	$1.1\,10^{-3}$(1)$^{[0.6,0.6]}$
	S_C	AD	91.6(1)	88.9(2)	$2.3\,10^{-4}$(1) $\downarrow$
		VG	92.37(6)$^{[0.6,1]}$	89.9(1)$^{[0.6,1]}$	$1.7\,10^{-4}$(1)$^{[1,0.6]}$
	S_S	AD	90.9(2)	88.1(3) $\uparrow$	$1.8\,10^{-4}$(2) $\downarrow$
		VG	88.4(1)$^{[1,0.6]}$	83.5(3)$^{[1,0.6]}$	$0.8\,10^{-4}$(1)$^{[1,0.6]}$

Table 7 Comparison between the VG and AD methods on BBCSport with top N% of features ranked by IG.

N	Sim	W	MFM(SD)	MAvG(SD)	AvD(SD)
50	S_d^p	AD	93.9(1)	92.4(2)	$1.0\,10^{-3}$(3)
		VG	93.8(6)$^{[1,1]}$	91.9(9)$^{[1,1]}$	$8.1\,10^{-4}$(9)$^{[0.6,0.6]}$
	S_e^p	AD	93.9(1)	92.4(2)	$1.0\,10^{-3}$(4)
		VG	93.8(6)$^{[1,1]}$	91.9(9)$^{[1,1]}$	$8.1\,10^{-4}$(9)$^{[0.6,0.6]}$
	S_m^p	AD	93.9(1)	92.4(2)	$0.5\,10^{-3}$(2)
		VG	93.8(6)$^{[1,1]}$	91.9(9)$^{[1,1]}$	$4.0\,10^{-4}$(4)$^{[0.6,0.6]}$
	S_z	AD	93.9(1)	92.4(2)	$1.0\,10^{-3}$(3)
		VG	93.8(6)$^{[1,1]}$	91.9(9)$^{[1,1]}$	$8.1\,10^{-4}$(9)$^{[0.6,0.6]}$
	S_n^p	AD	93.9(1)	92.4(2)	$2.1\,10^{-3}$(6)
		VG	93.8(6)$^{[1,1]}$	91.9(9)$^{[1,1]}$	$1.6\,10^{-3}$(1)$^{[0.6,0.6]}$
	S_C	AD	91.1(9)	88.1(1)	$3.4\,10^{-4}$(5) $\downarrow$
		VG	93.7(1)$^{[0.6,1]}$	89.0(1)$^{[0.6,1]}$	$2.6\,10^{-4}$(4)$^{[1,0.6]}$
	S_S	AD	90.4(1) $\uparrow$	87.2(2) $\uparrow$	$2.6\,10^{-4}$(3) $\downarrow$
		VG	87.6(1)$^{[1,0.6]}$	82.6(2)$^{[1,0.6]}$	$1.1\,10^{-4}$(1)$^{[1,0.6]}$
80	S_d^p	AD	93.9(1)	92.4(2)	$0.8\,10^{-3}$(2)
		VG	94.1(4)$^{[1,1]}$	92.2(6)$^{[1,1]}$	$5.8\,10^{-4}$(4)$^{[0.6,0.6]}$
	S_e^p	AD	93.9(1)	92.4(2)	$0.8\,10^{-3}$(2)
		VG	94.1(4)$^{[1,1]}$	92.2(6)$^{[1,1]}$	$5.8\,10^{-4}$(4)$^{[0.6,0.6]}$
	S_m^p	AD	93.9(1)	92.4(2)	$0.4\,10^{-3}$(1)
		VG	94.1(4)$^{[1,1]}$	92.4(6)$^{[1,1]}$	$2.9\,10^{-4}$(2)$^{[0.6,0.6]}$
	S_z	AD	93.9(1)	92.4(2)	$0.8\,10^{-3}$(2)
		VG	94.1(4)$^{[1,1]}$	92.2(6)$^{[1,1]}$	$5.8\,10^{-4}$(5)$^{[0.6,0.6]}$
	S_n^p	AD	93.9(1)	92.4(2)	$1.6\,10^{-3}$(5)
		VG	94.1(4)$^{[1,1]}$	92.2(2)$^{[1,1]}$	$1.1\,10^{-3}$(9)$^{[0.6,0.6]}$
	S_C	AD	91.6(1)	88.9(1)	$2.3\,10^{-4}$(1) $\downarrow$
		VG	92.4(6)$^{[0.6,1]}$	89.9(1)$^{[0.6,1]}$	$1.8\,10^{-4}$(1)$^{[1,0.6]}$
	S_S	AD	91.0(2)	88.3(3) $\uparrow$	$1.8\,10^{-4}$(2) $\downarrow$
		VG	88.4(1)$^{[1,0.6]}$	82.5(2)$^{[1,0.6]}$	$7.0\,10^{-5}$(7)$^{[1,0.6]}$

training data sets, and was chosen as a representative of k-NN. The 5-fold CV with the same separate datasets used in the previous section was used for evaluating the four models.

Table 8 compares our framework with the four classification models when used on BBC. The first two columns detail a method for calculating relevant scores of feature words and the numbers of selected features. The next column expresses a similarity measure or a classification model in use. The last two columns show evaluation results in terms of MFM and MAvG including standard deviation. Since all distance-based similarity measures give the same values of MFM and MAvG, only results of S_m^p are shown in this table. Note that, even if the memory was extended, Weka ran out of memory in the training process of SVM in both cases of selecting 30% of features. Hence there is no report for these cases. The table indicates that, regardless of similarity measures, our framework yields much higher performance than DT and k-NN. Using top 30% of features rating by GR, for example, S_m^p gives 95% and 94% of MFM and MAvG, respectively, while k-NN produces 52% and 44%. Comparing with NB and SVM, the table reveals that the performances of S_m^p and S_C are close to those of NB and SVM, while those of S_S are worse.

A paired t-test at $\alpha = 5\%$ was performed to assess the statistical significance of the difference between our framework with a particular measure and one of the four classifiers. The comparison result of $\uparrow$, $-$, or $\downarrow$ means that the measure is statistically superior, equal, or inferior to the classifier. The order of classifiers to be compared is DT, k-NN, NB, and SVM. For instance, the comparison result of MFM values obtained from S_C to those from the classifiers reported in the 15th row of Table 8 is $[\uparrow, \uparrow, -, \downarrow]$. It means that S_C is significantly better than DT and k-NN, equal to NB, and worse than SVM. Among the 4 cases of feature selection, S_m^p and S_C are statistically better than DT and k-NN when they gain the comparative performance to NB, and SVM, except the third case (i.e., 20% of features ranked by IG) where S_C is significantly worse than SVM. For S_S, it yields a higher accuracy than DT and k-NN, but lower than NB and SVM.

Likewise, Table 9 presents results evaluating BBCSport. It indicates that, no matter what simi-

Table 8 Comparison with other classifiers when evaluating BBC.

Feature selection	N	Sim/ Classifier	MFM(SD)	MAvG(SD)
GR	20	S_m^p	93.2(4)$^{[\uparrow,\uparrow,-,-]}$	92.6(4)$^{[\uparrow,\uparrow,-,-]}$
		S_C	92.8(1)$^{[\uparrow,\uparrow,-,-]}$	92.1(2)$^{[\uparrow,\uparrow,-,-]}$
		S_S	85.1(2)$^{[\uparrow,\uparrow,\downarrow,\downarrow]}$	82.4(2)$^{[\uparrow,\uparrow,\downarrow,\downarrow]}$
		DT	81.2(3)	81.0(3)
		k-NN	63.3(3)	58.2(2)
		NB	93.7(6)	93.8(6)
		SVM	93.4(7)	93.5(8)
GR	30	S_m^p	94.7(6)$^{[\uparrow,\uparrow,-]}$	94.4(7)$^{[\uparrow,\uparrow,-]}$
		S_C	93.4(1)$^{[\uparrow,\uparrow,-]}$	93.0(1)$^{[\uparrow,\uparrow,-]}$
		S_S	84.8(2)$^{[\uparrow,\uparrow,\downarrow]}$	82.8(2)$^{[\uparrow,\uparrow,\downarrow]}$
		DT	69.1(2)	68.4(2)
		k-NN	52.3(7)	44.0(6)
		NB	93.9(1)	94.0(1)
IG	20	S_m^p	94.4(9)$^{[\uparrow,\uparrow,-,-]}$	94.0(1)$^{[\uparrow,\uparrow,-,-]}$
		S_C	92.3(2)$^{[\uparrow,\uparrow,-,\downarrow]}$	91.8(2)$^{[\uparrow,\uparrow,-,\downarrow]}$
		S_S	82.7(2)$^{[\uparrow,\uparrow,\downarrow,\downarrow]}$	80.3(3)$^{[\uparrow,\uparrow,\downarrow,\downarrow]}$
		DT	69.8(3)	69.4(3)
		k-NN	60.2(4)	53.2(4)
		NB	93.9(9)	93.9(9)
		SVM	94.4(1)	94.5(2)
IG	30	S_m^p	94.5(9)$^{[\uparrow,\uparrow,-]}$	94.2(1)$^{[\uparrow,\uparrow,-]}$
		S_C	92.3(2)$^{[\uparrow,\uparrow,-]}$	91.8(2)$^{[\uparrow,\uparrow,-]}$
		S_S	83.6(2)$^{[\uparrow,\uparrow,\downarrow]}$	81.4(3)$^{[\uparrow,\uparrow,\downarrow]}$
		DT	66.5(3)	65.7(3)
		k-NN	49.8(3)	41.2(3)
		NB	94.1(8)	94.2(9)

Table 9 Comparison with other classifiers when evaluating BBCSport.

Feature selection	N	Sim/ Classifier	MFM(SD)	MAvG(SD)
GR	50	S_m^p	93.9(1)$^{[\uparrow,\uparrow,-,\uparrow]}$	92.5(2)$^{[\uparrow,\uparrow,\downarrow,-]}$
		S_C	91.3(6)$^{[\uparrow,\uparrow,\downarrow,-]}$	88.4(1)$^{[\uparrow,\uparrow,\downarrow,-]}$
		S_S	90.2(1)$^{[\uparrow,\uparrow,\downarrow,-]}$	87.2(2)$^{[\uparrow,\uparrow,\downarrow,\downarrow]}$
		DT	70.4(4)	69.9(3)
		k-NN	41.5(6)	30.1(6)
		NB	96.1(1)	96.2(1)
		SVM	89.8(4)	91.3(3)
GR	80	S_m^p	93.7(1)$^{[\uparrow,\uparrow,-,-]}$	92.2(2)$^{[\uparrow,\uparrow,-,-]}$
		S_C	91.6(1)$^{[\uparrow,\uparrow,\downarrow,-]}$	88.9(2)$^{[\uparrow,\uparrow,\downarrow,-]}$
		S_S	90.9(2)$^{[\uparrow,\uparrow,\downarrow,-]}$	88.1(3)$^{[\uparrow,\uparrow,-,-]}$
		DT	63.4(4)	61.6(4)
		k-NN	41.8(6)	30.6(7)
		NB	95.3(2)	95.7(2)
		SVM	89.8(45)	91.0(4)
IG	50	S_m^p	93.9(1)$^{[\uparrow,\uparrow,-,-]}$	92.4(2)$^{[\uparrow,\uparrow,\downarrow,-]}$
		S_C	91.1(9)$^{[\uparrow,\uparrow,\downarrow,-]}$	88.1(1)$^{[\uparrow,\uparrow,\downarrow,-]}$
		S_S	90.4(1)$^{[\uparrow,\uparrow,\downarrow,-]}$	87.2(2)$^{[\uparrow,\uparrow,\downarrow,-]}$
		DT	69.9(4)	68.9(4)
		k-NN	42.2(9)	30.2(9)
		NB	95.9(1)	96.3(2)
		SVM	89.4(4)	90.9(4)
IG	80	S_m^p	93.9(1)$^{[\uparrow,\uparrow,-,-]}$	92.4(2)$^{[\uparrow,\uparrow,-,-]}$
		S_C	91.6(1)$^{[\uparrow,\uparrow,\downarrow,-]}$	88.9(2)$^{[\uparrow,\uparrow,\downarrow,-]}$
		S_S	91.0(2)$^{[\uparrow,\uparrow,\downarrow,-]}$	88.3(3)$^{[\uparrow,\uparrow,\downarrow,\downarrow]}$
		DT	66.6(5)	65.5(5)
		k-NN	40.4(2)	28.6(1)
		NB	95.5(2)	95.8(2)
		SVM	90.1(4)	91.2(4)

larity measure was used, the proposed framework clearly outperforms DT and k-NN. S_m^p shows comparable performance to NB and SVM. S_C is worse than NB, while S_S is worse than both NB and SVM.

CONCLUSIONS

With statistical concepts, documents can be represented in terms of IFSs and patterns for predefined document classes can be constructed. A similarity measure is used to determine a similarity degree between an IFS for a document and a class pattern. The document is grouped into the class such that its pattern is closest to the IFS for the document. Using some existing similarity measures for IFSs, the experiment on two datasets shows that our framework yields satisfactory results when compared to results from traditional classification models. Further work includes extension of the dataset, investigation of various techniques for IFS-based document representation, and in-depth analysis of IFS to facilitate text classification.

Acknowledgements: The author gratefully acknowledges the financial support provided by Thammasat University Research Fund under the TU Research Scholar scheme, Contract No. 1/8/2557.

REFERENCES

1. Zadeh LA (1965) Fuzzy sets. *Inform Contr* **8**, 338–53.
2. Atanassov K (1986) Intuitionistic fuzzy sets. *Fuzzy Set Syst* **20**, 87–96.
3. Dengfeng L, Chuntian C (2002) New similarity measures of intuitionistic fuzzy sets and application to pattern recognition. *Pattern Recogn Lett* **23**, 221–5.
4. Liang Z, Shi P (2003) Similarity measures on intuitionistic fuzzy sets. *Pattern Recogn Lett* **24**, 2687–93.
5. Mitchell HB (2003) On the Dengfeng-Chuntian similarity measure and its application to pattern recognition. *Pattern Recogn Lett* **24**, 3101–4.
6. Hung WL, Yang MS (2004) Similarity measures of intuitionistic fuzzy sets based on Hausdorff distance. *Pattern Recogn Lett* **25**, 1603–11.

7. Xu Z (2007) Some similarity measures of intuitionistic fuzzy sets and their applications to multiple attribute decision making. *Fuzzy Optim Decis Making* **6**, 109–21.
8. Khatibi V, Montazer GA (2009) Intuitionistic fuzzy set vs. fuzzy set application in medical pattern recognition. *Artif Intell Med* **47**, 43–52.
9. Ye J (2011) Cosine similarity measures for intuitionistic fuzzy sets and their applications. *Math Comput Model* **53**, 91–7.
10. Hwang CM, Yang MS (2012) Modified cosine similarity measure between intuitionistic fuzzy sets. In: 4th International Conference, AICI 2012, Chengdu, China, pp 285–93.
11. Li Y, Olson D, Qin Z (2007) Similarity measures between intuitionistic fuzzy (vague) sets: A comparative analysis. *Pattern Recogn Lett* **28**, 278–85.
12. Papakostas GA, Hatzimichailidis AG, Kaburlasos VG (2013) Distance and similarity measures between intuitionistic fuzzy sets: A comparative analysis from a pattern recognition point of view. *Pattern Recogn Lett* **34**, 1609–22.
13. Sebastiani F (2002) Machine learning in automated text categorization. *ACM Comput Surv* **34**, 1–47.
14. Khan A, Baharudin B, Lee LH, Khan K (2010) A review of machine learning algorithms for text-documents classification. *J Adv Inform Tech* **1**, 4–20.
15. Mita KD, Mukesh AZ (2012) Automatic text classification: a technical review. *Int J Comput Sci Appl* **28**, 37–40.
16. Szmidt E, Kacprzyk J (2008) Using Intuitionistic Fuzzy Sets in Text Categorization. In: *9th International Conference on Artificial Intelligence and Soft Computing*, Zakopane, Poland, pp 351–62.
17. Julian P, Hung KC, Lin SJ (2012) On the Mitchell similarity measure and its application to pattern recognition. *Pattern Recogn Lett* **33**, 1219–23.
18. Greene D, Cunningham P (2005) Producing accurate interpretable clusters from high-dimensional data. In: *9th European Conference on Principles and Practice of Knowledge Discovery in Databases*, pp 486–94.
19. Sun Y, Kamel MS, Wang Y (2006) Boosting for learning multiple classes with imbalanced class distribution. In: *6th International Conference on Data Mining*, pp 592–602.