An Automated Error Detection for News Webpages of Chinese Portal

Deng-Yiv Chiu and Chi-Chung Lee
Chung Hua University / Department of Information Management, Hsin-Chu, Taiwan, ROC
Email: {chiuden, leecc}@chu.edu.tw

Ya-Chen Pan
Yuanpei University / Department of Information Management, Hsin-Chu, Taiwan, ROC
Email: ayako@mail.ypu.edu.tw

***Abstract*—There exists some news obviously classified into incorrect categories on Chinese webpage portals. This phenomenon is owing mainly the difficulty in automatically classifying Chinese news and the fact that news appearing on webpage portals is retrieved from numerous media sources. This study integrates genetic algorithms and multi-class support vector machine classifiers to construct an automated classification error detection approach for Chinese news classification. A genetic algorithm is utilized to select four feature thresholds used to obtain representative features/words of each class. The multi-class SVM classifier is then trained to construct an appropriate classifier to aid automated classification error detection. The experiment applies the proposed method to the Chinese news on Taiwan Yahoo!**

***Index Terms*—multi-class support vector machine, genetic algorithm, news classification error detection**

I. INTRODUCTION

With the rapid development of the Internet, a growing number of people are browsing news webpages via online portals. Consequently, how to automatically classify an enormous amount of news retrieved from numerous media sources efficiently and correctly becomes very important.

Various methods have been widely applied to document classification. For example, decision tree is commonly applied to classify webpages. It not only classifies documents rapidly, but also transforms the classification result into a logical relationship that can be easily understood. Based on the value of tree node, decision tree method classifies documents for the next layer with a tree hierarchy [1] [2]. The structure of the decision tree can be understood easily, but cannot be used to improve the results of complex classification problems.

The KNN algorithm has been applied to incorporate the relationships of concept-based thesauri into the document categorization conducted on electronic-product review directories in Yahoo Korean web site [3]. It has also been applied to classify documents in the Reuter and TDT2 corpus and to deal with unbalanced document categorization problems [4].

Some scholars have applied multi-class SVM to explore the relationship among classes for multi-class problems. For example, hierarchical SVM has been applied to classify multi-class documents using the support vector clustering method and mirror the class hierarchy conducted on Reuter document [5].

Fuzzy SVM (FSVM) has been applied to classify multi-class documents based on fuzzy set theory and OAA-SVM conducted on Reuter documents [6]. The main advantage of FSVM is that it can solve the problem of one document belonging to multiple classes through membership functions.

This study investigates the classification problem of Chinese-portal news webpages retrieved from numerous media sources. The news webpages of Taiwan Yahoo! are used as the research targets. This study integrates genetic algorithms and multi-class support vector machine classifiers to construct an automated classification error detection approach for Chinese news classification. A genetic algorithm is utilized to select four feature thresholds to obtain representative features/words of each class used to construct the vector space model for each document. The multi-class SVM classifier is then trained to construct an appropriate classifier to aid automated classification error detection.

II. MULTI-CLASS SUPPORT VECTOR MACHINE

As an efficient learning machine, support vector machine (SVM) has been employed to solve the binary classification problem. However, its applicability to multi-class problems remains the subject of ongoing research. Here, we introduce two multi-class SVM classifiers used to solve multi-class problems, including one-against-all SVM and one-against-one SVM.

A. One-against-all SVM

The one-against-all SVM strategy was proposed in 1994 and was employed to solve the multi-class classification problems in some researches [7]. For the

Corresponding author: Ya-Chen Pan
Email: ayako@mail.ypu.edu.tw

doi:10.4304/jsw.5.12.1334-1341

classification problem involving *k* classes, *k* binary SVM classifiers are established to optimize hyper-plane separation for each class. A binary classifier SVMj is trained using training samples with all samples belonging to class *Cj* as positive samples and all samples not belonging to class *Cj* as negative samples. To classify new input with trained SVMs, the input is sent to *k* SVM classifiers for evaluation. The "Winner-take-all" rule is adopted to determine the class to which the new input should be assigned. The class owning the highest decision function value is selected as the class of the new input.

In training process, hyper-plane optimizes only one class C_j because samples in other classes are regarded as opposite samples. Given l training samples $(\mathbf{x}_1,y_1),(\mathbf{x}_2,y_2),...,(\mathbf{x}_l,y_l)$, where $\mathbf{x}_j \in R^d$, $j=1,...,l$ and $y_j \in \{1,...,k\}$ is the class of $\mathbf{x}_j$, the *i*th SVM solves the following optimization problem:

$$min\frac{1}{2}\left\|\mathbf{w}^i\right\|^2 + C\sum_{i=1}^{l}\xi_j^i .$$

Constraint:

$$\left(\mathbf{w}^j \cdot \Phi\left(\mathbf{x}_j\right)\right) + b^i \geq 1-\xi_j^i,\ if\ y_j = i .$$

$$\left(\mathbf{w}^j \cdot \Phi\left(\mathbf{x}_j\right)\right) + b^i \leq -1+\xi_j^i,\ if\ y_j \neq i .$$

$$\xi_j^i \geq 0,\ \ j=1,...,l .$$

where
w is weight vector
C is a constant
ξ is slash variable
b is bias
Φ is kernel function.

And decision function $f_i(\mathbf{x})$ is defined as in Eg.(1):

$$f_i(\mathbf{x}) = \left(w^i \cdot \Phi(\mathbf{x})\right) + b^i . \quad (1)$$

The classification result of sample **x** could be defined as in Eq. (2):

$$\mathbf{x} = arg\ max_{i=1,...,k}\left(w^i \cdot \Phi(\mathbf{x}) + b^i\right). \quad (2)$$

One-against-all SVM has two characteristics. First, it uses few binary SVM classifiers. And, the count of positive samples is much less than the count of negative samples since position samples are from one single class but negative samples are from all other classes.

B. One-against-one SVM

The one-against-one SVM strategy employs a binary SVM classifier for any pair of two classes [8]. That is, for the classification problem with *k* classes, *k* (*k* -1)/2 binary SVM classifiers are built with training data. To classify a new input with trained SVM, the input is sent to *k*(*k* -1)/2 SVM classifiers for voting. The "Max wins" rule is adopted and voting is employed to determine the class to which the new input should be classified. The class with the maximum of votes is the winner.

One-against-one SVM would find the hyper-plane between classes. It is found by solving the following optimization problem:

$$min\frac{1}{2}\left\|\mathbf{w}^{ij}\right\|^2 + C\sum_n \xi_n^{ij}$$

Constraint:

$$\left(\mathbf{w}^{ij} \cdot \Phi\left(\mathbf{x}_n\right)\right) + b^{ij} \geq 1-\xi_n^{ij},\ if\ y_n = i$$

$$\left(\mathbf{w}^{ij} \cdot \Phi\left(\mathbf{x}_n\right)\right) + b^{ij} \geq -1+\xi_n^{ij},\ if\ y_n = j$$

$$\xi_i^{nm} \geq 0,\ i \neq j, i, j \in \{1,...,k\}$$

for all *n* examples in class *i* and *j*
where
w is weight vector
C is a constant
ξ is slash variable
b is bias
Φ is kernel function.

The decision function for class pair *ij* is defined in Eq. (3).

$$f_{ij}(\mathbf{x}) = \left\langle w^{ij} \cdot \Phi(\mathbf{x}) \right\rangle + b^{ij} . \quad (3)$$

The classification result of sample **x** could be defined as in Eq. (4):

$$\mathbf{x} = arg\ max_i \sum_{j\neq i, j=1}^{k} \text{sign}\left(f_{ij}(\mathbf{x})\right). \quad (4)$$

If more than one class have the highest number of votes (a tie). The decision function value is used to determine the classification result of sample x, as expressed in Eq. (5):

$$\mathbf{x} = arg\ max_i \sum_{j\neq i, j=1}^{k} f_{ij}(\mathbf{x}). \quad (5)$$

One-against-one SVM has two characteristics. First, it uses more binary SVM classifiers. Secondly, the count of positive samples is not much less than that of negative samples since they are from one single class, respectively. The disadvantage of this method is that the number of classifiers rises rapidly with the number of classes, lengthening the time taken to make the decision.

III. THE PROPOSED APPROACH

This study employs genetic algorithm and multi-class SVM classifiers to present an automated error detection approach for Chinese news classification of web portal.

A. Classification with GA and multi-class SVM methods

The purpose of multi-class GA-SVM method is to construct multi-class SVM classifiers using a combination of genetic algorithm and multi-class support vector machine. In this method, first of all, we collect the Chinese news documents from Taiwan Yahoo. The collected documents are segmented by the Chinese

Words Database and CKIP Chinese Word Segmentation System. According to the morphological features, lexicons are divided into several word classes and the necessary word classes are captured. Four feature threshold values are calculated to find representative features/keywords of documents [9]. In order to extract truly representative features, genetic algorithm with optimal parameter character is used to determine the threshold values. Then the vector space models for training documents are established with features having related values greater than four thresholds.

In the process of genetic algorithm, the first generation of four thresholds is initialized at random. Then genes are decoded to facilitate the settings for four thresholds in this generation and representative features satisfying the feature thresholds will be selected. The selected features are used to build the vector space model (VSM) of each news document and the VSMs are used to train multi-class SVM classifiers.

In order to get better classification performance, one-against-one SVM method is used (because the performance of one-against-one SVM is better than one-against-all SVM in our experiments). Fitness function of genetic algorithm is used to evaluate the classification performance. The higher fitness function value means the better classification performance. Evolution is performed until 100 generations. The genetic algorithm will be performed until this criterion is met, so as to obtain four optimal thresholds. Then, training data is used to train the multi-class SVM classifier. Finally, Chinese news documents used as testing data could be classified based on the trained classifiers.

(a) Representative feature selection for each class

The purpose of feature selection is to obtain representative feature set and reduce noise in a specific field. In the study, four thresholds including term frequency, document frequency, uniformity and conformity are used for selecting representative features [9].

(1) Term Frequency

Term frequency (TF) denotes the weight of occurrence probability of feature t_i in class C_j, as expressed in Eq. (6).

$$TF_{ij} = \frac{T\hat{F}_{ij}}{\sum_{m=1}^{J} T\hat{F}_{im}}. \quad (6)$$

$$T\hat{F}_{ij} = \frac{t_{ij}}{\sum_{n=1}^{I} t_{nj}}. \quad (7)$$

I the number of features
J the number of classes
t_{ij} the count of occurrences of feature t_i in class C_j
$T\hat{F}_{ij}$ the occurrence probability of feature t_i in class C_j

The feature with higher TF value means that the feature can represent the class better.

(2) Document Frequency

Document Frequency (DF) denotes the weight of occurrence probability of documents with feature t_i in class C_j, as expressed in Eq. (8).

$$DF_{ij} = \frac{D\hat{F}_{ij}}{\sum_{m=1}^{J} D\hat{F}_{im}}. \quad (8)$$

$$D\hat{F}_{ij} = \frac{l_{ij}}{l_j}. \quad (9)$$

l_{ij} the count of documents with feature t_i in class C_j
l_j the total count of all documents in class C_j
$D\hat{F}_{ij}$ occurrence probability of documents with feature t_i in class C_j

The feature with higher DF values can represent the class better since it appears more frequently in documents in the class than in documents in other classes.

(3) Uniformity

Uniformity denotes the occurrence weight of feature t_i appearing in all documents in class C_j, as expressed in Eq. (10).

$$U_{ij} = -\sum_{k=1}^{l_j} q_{ik} \log q_{ik}. \quad (10)$$

$$q_{ik} = \frac{tf_{ik}}{\sum_{m=1}^{l_j} tf_{im}}. \quad (11)$$

l_j the total count of all documents in class C_j
tf_{ik} the occurrence times of feature t_i appearing in document d_k in class C_j
q_{ik} the occurrence weight of feature t_i appearing in document d_k in class C_j

The feature with higher uniformity value can represent the class better than other features since the feature appears more frequently in class C_j than other features.

(4) Conformity

Conformity denotes the occurrence weight of documents with feature t_i appearing in all classes, as expressed in Eq. (12).

$$CF_i = -\sum_{j=1}^{J} d_{ij} \log d_{ij}. \quad (12)$$

$$d_{ij} = \frac{l_{ij}}{\sum_{m=1}^{J} l_{im}}. \quad (13)$$

d_{ij} the occurrence weight of documents in class C_j with feature t_i appearing in all classes

l_{ij} the count of documents with feature t_i in class C_j

The feature with smaller conformity value can represent the class better since the feature appears in less classes.

(b) Fitness function of GA-SVM approach

In this section, in order to select representative features of each class, we design the proper fitness function used in genetic algorithm to determine four feature thresholds. The fitness function for feature selection of Chinese news in this research mainly considers three important factors, precision, recall, and F-measure.

Precision denotes the percentage of count of documents classified correctly into a class to count of documents classified into the class. Recall denotes the percentage of count of documents classified correctly into a class to count of documents belonging to the class. The equations of precision and recall are expressed as below.

$$P_{s,p,C_i} = \frac{N_{p,C_i} \bigcap N_{s,C_i}}{N_{s,C_i}}. \quad (14)$$

$$R_{s,p,C_i} = \frac{N_{p,C_i} \bigcap N_{s,C_i}}{N_{p,C_i}}. \quad (15)$$

C_i class C_i

N_{s,C_i} total count of documents classified into class C_i by certain method.

N_{p,C_i} total count of documents belonging to C_i.

$N_{p,C_i} \bigcap N_{s,C_i}$ total count of documents classified correctly to class C_i by certain method.

F-measure value is computed to consider both of precision and recall simultaneously. If the values of precision and recall are higher, the value of F-measure is higher. The equation of F-measure is expressed in Eq. (16).

$$F_{s,p,C_i} = \frac{2P_{s,p,C_i} * R_{s,p,C_i}}{P_{s,p,C_i} + R_{s,p,C_i}}. \quad (16)$$

Therefore, when the values of precision, recall and F-measure of a chromosome are higher, the value of fitness function should be better and the classification performance with selected features should be better.

Therefore, the fitness function for the proposed multi-class GA-SVM approach is shown in Eq. (17).

$$Fitness\ \ Function = \frac{\sum_{i=1}^{I} F_{C_i}}{I}. \quad (17)$$

where I is the number of classes

F_{C_i} is the F-measure value of class C_i.

B. The architecture of the proposed method

The architecture is as shown in Fig. 1. The detailed explanation is as follows.

(1) *Data collection*: This study collects documents as data from the news webpages of Taiwan, Yahoo! http://tw.yahoo.com. The experimental data comprises electronic documents of Chinese news collected from 31 different media sources.

(2) *Word segmentation with CKIP:* The collected documents are segmented using the CKIP Chinese Word Segmentation System. Lexicons are divided into several word classes according to the morphological features.

(3) *Selection of feature candidates:* In order to get representative features, features belonging to general noun (Na), place noun (Nc) and terminology (Nb) are selected to form candidate features. Additionally, this study also deletes the candidate features with the string length of one to reduce document noise.

(4) *Initialization of feature selection thresholds for GA process*: The first generation of four thresholds is initialized at random. A generation includes 20 chromosomes, each of which consists of the four thresholds used to select representative features.

(5) *Converting genotype to phenotype:* Genes are decoded to help set the four thresholds.

(6) *Presenting training documents with vector space model (VSM):* To form the vector space model of each document, this study selects features of each class that satisfy the four thresholds.

(7) *Training SVM classifier with VSMs of training documents:* The vector space models of the training data are used to train the SVM classifier. The one-against-one multi-class SVM classifier is employed in the experiment.

(8) *Evaluation of fitness values of classifier*: The fitness value of the classification performance is calculated. Larger value indicates higher classification performance.

(9) *Termination criterion of genetic algorithm for classifier training*: The termination criterion is evolution of 100 generations. If the criterion is met, then go to step (11).

(10) *Process of genetic algorithm*: Chromosome evolution in a genetic algorithm involves three processes, namely selection, crossover, and mutation. In the selection process, the 10 chromosomes with the highest fitness values are selected and duplicated by roulette wheel selection. The double-point crossover method is utilized. The mutation rate is defined as 1%, and the process is redirected to step (5).

(11) *Classification of testing documents with trained SVM classifier*: After obtaining the optimal SVM trained model, testing documents are sent to the trained classifiers for classification.

IV. EXPERIMENTS

We introduce empirical data, the empirical process and results of proposed GA-SVM, and comparison with other methods.

A. Empirical data

Experimental data and existing document classification structure from yahoo.com.tw in August, 2008 are collected. The document classification structure comprises six classes, including Politics, Finance, Health, Education, Sports, and Film. Table I lists the distribution of collected documents and class titles. The table contains a total of six classes and 5657 documents, of which 4469 are used as training data for SVM classifier and 1188 documents are used for testing. The testing data comprises both news classified correctly and incorrectly originally.

B. Empirical process of GA-SVM classifier

A genetic algorithm is utilized to select feature thresholds to retrieve representative features. Table II shows the chromosomes with the top five fitness function values. The number of features for the best chromosome is 7269, satisfying the four thresholds. The best fitness value is 0.9168.

The trained GA-SVM classifier is utilized to verify the

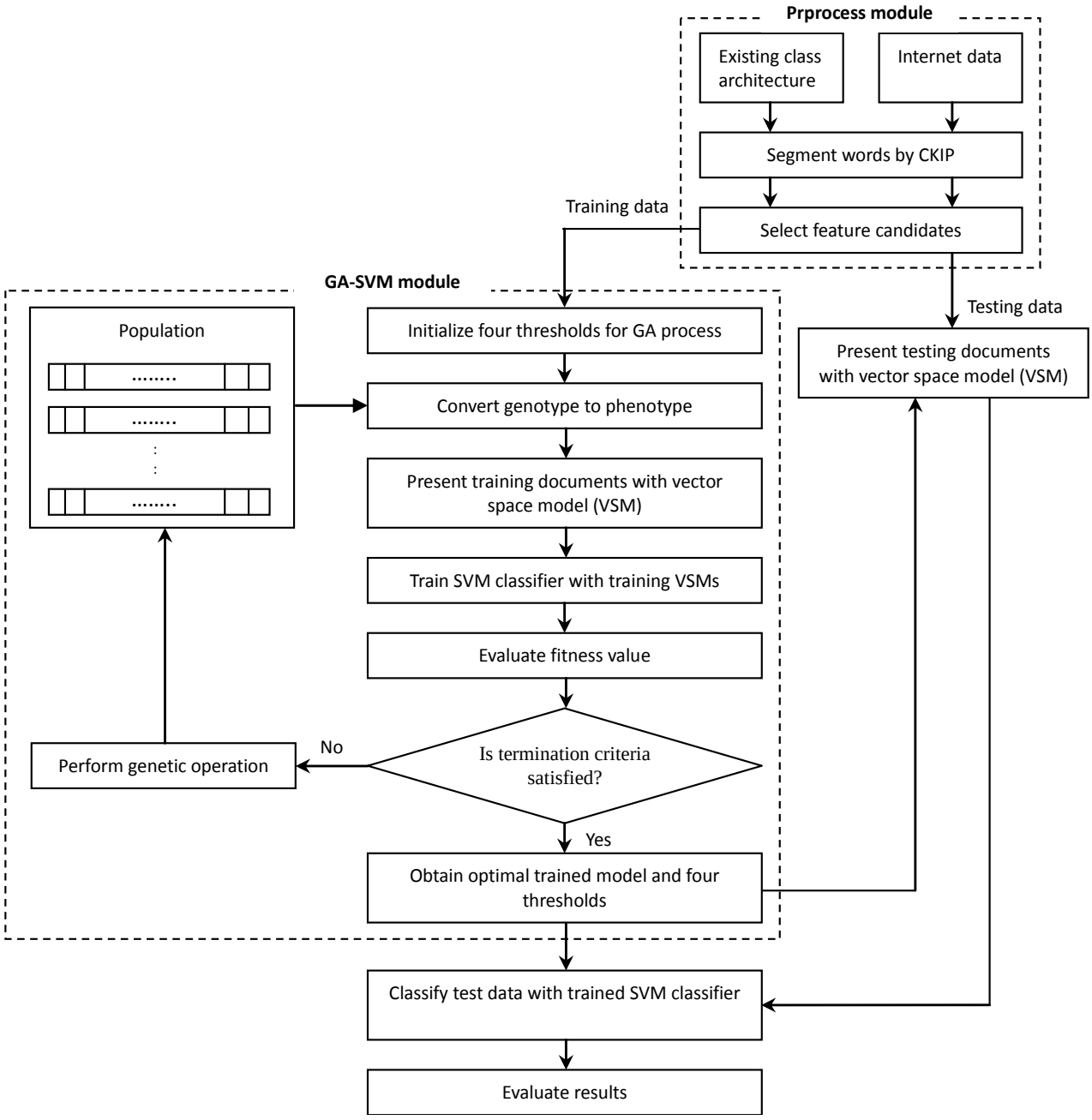


FIGURE 1. The architecture of proposed approach

classification performance of Chinese news. Table III shows the precision, recall and F-measure. The average precision rate is 83.68%, the average recall is 80.02%, and the average F-measure is 81.47% for the proposed approach.

To verify the detection ability of the trained GA-SVM classifier, we collect 321 originally incorrectly classified news during the same period as shown in Table IV. The detection rate of news originally classified incorrectly is 88.79% by the proposed approach (see Table IV). Each class reaches excellent result except class Politics. The main reason is that many features of those Politics news are also the features of news in other classes. We can overcome the low detection rate of class Politics by adding manually politician's name as features with better term frequency, document frequency, uniformly and conformity.

Some examples of originally classified incorrectly and detected by the proposed approach are shown in Table V. We can see that those news are obviously classified into incorrect category.

We can see that the proposed approach can find out incorrectly classified Chinese news efficiently. Experimental results reveal that the trained classifier has excellent performance.

TABLE I.
DATASETS FROM TAIWAN YAHOO! WEB SITE

No.	Class title	Training data	Testing data
1	Politics	996	279
2	Finance	1025	286
3	Health	342	64
4	Education	417	91
5	Sports	1027	291
6	Film	662	177
	Sum	4469	1188

C. Comparisions

To verify the performance of the proposed approach, we also perform experiments by employing SVM classifier and KNN (K=7) classifier with the same empirical data as shown in Table III and Fig. 2. The F-measure of the proposed approach outperforms the SVM and KNN methods by 4.81% and 17.14%, respectively. The result of proposed approach is valuable reference for the problem of automated Chinese news classification. One of main reasons is that the feature thresholds in GA process have selected representative features for each class to form the vector space model for each document. Also, the multi-class SVM provides the property of significant classification ability for nonlinearity and high dimensionality problem.

V. CONCLUSIONS

This study proposes an appropriate classification approach to assist error detection of Chinese news

TABLE II.
CHROMOSOMES WITH TOP FIVE FITNESS FUCTION VALUES PRODUCED IN TRAINING PROCESS

Chromosome	Term Frequency	Document Frequency	Uniformity	Conformity	Number of Features	Fitness Value
1	0.0314	0.0118	0.2784	0.4941	7269	0.91618
2	0.0784	0.2902	0.2862	0.4863	7129	0.91537
3	0.1412	0.2510	0.0706	0.4980	8379	0.91405
4	0.2824	0.0353	0.2471	0.4863	7914	0.91396
5	0.3255	0.1020	0.0510	0.4980	8369	0.91391

TABLE III.
THE PERFORMANCE OF THE PROPOSED GA-SVM APPROACH, SVM AND KNN CLASSIFIERS

Class title	Proposed GA-SVM classifier			SVM classifier			KNN classifier		
	Precision	Recall	F-measure	Precision	Recall	F-measure	Precision	Recall	F-measure
Politics	90.61%	79.57%	84.73%	81.21%	82.08%	84.64%	98.15%	38.00%	54.78%
Finance	76.65%	89.51%	82.58%	54.36%	93.71%	68.81%	42.19%	94.41%	58.32%
Health	83.93%	73.44%	78.33%	84.75%	78.13%	81.30%	71.64%	75.00%	73.28%
Education	72.15%	62.64%	67.06%	60.66%	81.32%	69.48%	36.67%	84.62%	51.16%
Sports	86.65%	95.88%	91.03%	84.19%	95.19%	89.35%	88.81%	81.79%	85.15%
Film	92.11%	79.10%	85.11%	57.44%	78.53%	66.35%	55.56%	73.45%	63.26%
Average	83.68%	80.02%	81.47%	70.43%	84.82%	76.66%	65.50%	74.54%	64.33%

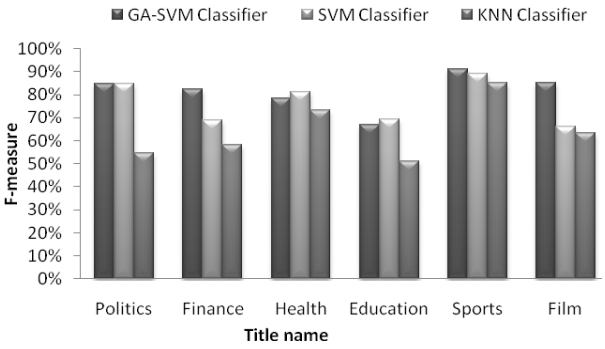


Figure 2. Comparison of the proposed approach, SVM and KNN classifiers.

TABLE IV.
THE DETECTION RATE OF PROPOSED GA-SVM APPROACH

Class title	Number of News Classified Incorrectly Originally	Number of News Detected	Detection rate
Politics	33	13	39.39%
Finance	36	36	100.00%
Health	14	14	100.00%
Education	83	80	96.39%
Sports	49	36	73.47%
Film	106	106	100.00%
Average	53.5	47.5	88.79%

TABLE V.
EXAMPLES OF NEWS ORIGINALLY CLASSIFIED INCORRECTLY AT TAIWAN YAHOO! WEB SITE

Date	The Title of News	The Class of News Originally Classified	The Correct Class of the News
8/13	An 78-year-old man completed his Bachelor's degree.	Politics	Education
8/15	Kaohsiung summer math camps come to the end.	Politics	Education
8/14	Frog king Kosuke Kitajima won the swimming world championship.	Finance	Sports
8/17	The softball competition is excluded from the Olympic.	Finance	Sports
8/13	The elementary school students should watch out for strange taxi drivers.	Health	Education
8/16	Pre-president Chen did not show out in public for long time.	Education	Politics
8/17	An Egypt woman gave birth to seven babies, four boys and three girls.	Education	Health
8/14	The lawyers of pre-president Chen defended as his client.	Sports	Politics
8/14	To control the quality of high school teaching, the repeater rule is restored.	Sports	Education
8/13	The hospital announced that councilman Yen, Ching-Piao was ill.	Film	Politics

classification based on information retrieval, multi-class support vector machine, and genetic algorithm. The proposed method is tested on Taiwan Yahoo News webpages.

Research results reveal that the automated classification error detection approach constructed by multi-class GA-SVM helps automatically detect Chinese news classification efficiently.

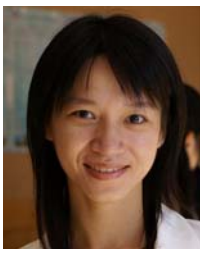
Future studies can incorporate the proposed method into other semantic analysis method to retrieve representative features. Additionally, the kernel function selection and parameter setting for SVM can also be explored to improve empirical performance.

REFERENCES

[1] T. M. Mitchell, *Machine Learning*. McGraw-Hill, 1997.
[2] J. R. Quinlan, "Decision trees as probabilistic classifiers," *International Workshop on Machine Learning*, 1987, pp. 31–37.
[3] S. L. Bang, J. D. Yang, and H. J. Yang, "Hierarchical document categorization with K-NN and concept-based thesauri," *Information Processing and Management*, vol.42 ,no. 2, 2006, pp. 387–406.
[4] S. Tan, "Neighbor-weighted K-Nearest neighbor for unbalanced text corpus," *Expert Systems with Applications*, 2005, vol. 28, no. 4, pp. 667-671.
[5] P. Y. Hao, J. H. Chiang, and Y. K. Tu, "Hierarchically SVM classification based on support vector clustering method and its application to document categorization," *Expert Systems with Applications*, 2007, vol. 33, no. 3, pp. 627–635.
[6] T. Y. Wang, and H. M. Chiang, "Fuzzy support vector machine for multi-class text categorization," *Information Processing & Management*, 2007, vol. 43, no .4, pp. 914–929.
[7] L. Bottou, C. Cortes, J. Denker, H. Drucker, I. Guyon, L. Jackel, Y. LeCun, U. Muller, E. Sackinger, P. Simard, and V. Vapnik, "Comparison of classifier methods: A case study in handwriting digit recognition," *International Conference on Pattern Recognition*, 1994, pp. 77–87.
[8] S. Knerr, L. Personnaz, and G. Dreyfus, "Single-layer learning revisited: a stepwise procedure for building and training a neural network," *Neurocomputing: Algorithms, Architectures and Applicatio*, F68, Springer-Verlag, 1990, pp. 41–50.
[9] C. H. Chou, C. C. Han, and Y. H. Chen, "GA based optimal keyword extraction in an automatic Chinese web document classification system," *Lecture notes in Computer Science*, 2007, vol.4743, pp. 224–234.

Deng-Yiv Chiu received the B.A. from Averett College, Virginia, USA in 1988, M.S. from University of Maryland, USA in 1990. He received the Ph.D. in Computer Science from Illinois Institute of Technology, USA in 1994. After working as an assistant professor at Dept. of Math and Computer Science, Chicago State University, USA, he has been an associate professor/ full professor at Chung Hua University, HsinChu, Taiwan since 1996. His research interests include machine learning, information retrieval, and their applications to knowledge management and finance.

Chi-Chung Lee received B.S. in Industrial Engineering from Chung Yuan Christian University, Taoyuan, Taiwan in 1991. Then, he received the M.B.A. and Ph.D. in Information Management from National Taiwan University of Science and Technology, Taipei, Taiwan in 1993 and 2004, respectively. He is an assistant professor at Dept. of Information Management, Chung Hua University, Taiwan. His current research interests include database systems, data modeling, and data management for mobile computing.



Ya-Chen Pan received the B.A., M.A., and Ph.D from Chung Hua University, Taiwan in 2003, 2005 and 2009, respectively. She is a teacher at Dept. of Management Information System, Yuanpei University, Taiwan. She was born in Yunlin, Taiwan in 1981. Her study interests include information retrieval, knowledge management, and machine learning.