Xiaoyong LIU, Hui FU

Guangdong Polytechnic Normal University (1)

A Hybrid Algorithm for Text Classification Problem

Abstract. *This paper investigates a novel algorithm-EGA-SVM for text classification problem by combining support vector machines (SVM) with elitist genetic algorithm (GA). The new algorithm uses EGA, which is based on elite survival strategy, to optimize the parameters of SVM. Iris dataset and one hundred pieces of news reports in Chinese news are chosen to compare EGA-SVM, GA-SVM and traditional SVM. The results of numerical experiments show that EGA-SVM can improve classification performance effectively than the other algorithms. This text classification algorithm can be extended easily to apply to literatures in the field of electrical engineering.*

Streszczenie. *W artykule przedstawiono nowy algorytm klasyfikacji tekstu bazujący na mechanizmie SVM (support vector machine) I algorytmie genetycznym. Algorytm zbadano na podstawie bazy danych Iris i setek innych chińskich przykładów. Algorytm wykazał swoją skuteczność. Może być on łatwo rozszerzony na analizę tekstów w inżynierii elektrycznej (**Hybrydowy algorytm do klasyfikacji tekstu**)*

Keywords: Genetic Algorithm, support vector machines, Text Classification
Słowa kluczowe: algorytm genetyczny, SVM, klasyfikacja tekstu

1. Introduction

Netnews and electrical literatures are more and more along with rapid popularization of internet and sharp development of information technology [1]. Because Internet is as the principal source of information access by users, how to access valid information from lots of web pages becomes a popular problem that is worth to studying. There are two directions to classify web pages. In the case of methodology, algorithms of text classification are often used to classify web pages according to pages' content, such as, Rough Sets[2], Bayes [3], SVM[4], etc. This paper investigates a novel algorithm that combines elitist genetic algorithm and support vectors machine and takes web pages as an example to apply to text classification problem.

The structure of this paper is as following, section 2 introduces related basic concepts, such as, SVM, GA, etc.. Section 3 introduces the novel algorithm-EGA-SVM, which is presented in this study. Section 4 gives results of different models in two datasets, one is iris dataset from University of California Irvine Machine Learning Repository, and the other is web pages dataset. Finally, conclusions and future research consideration are presented in Section 5.

2. Basic concepts of models

Support vector machines[5][6]

Support vector machines (SVM) are a set of related supervised learning methods used for classification and regression. A support vector machine constructs a hyper plane or set of hyper planes in a high-dimensional space, which can be used for classification, regression or other tasks. Intuitively, a good separation is achieved by the hyper plane that has the largest distance to the nearest training data points of any class (so-called functional margin), since in general the larger the margin the lower the generalization error of the classifier.

In order to extend the SVM methodology to handle data that is not fully linearly separable, we relax the constraints slightly to allow for misclassified points, the formulation is following (1.1) and (1.2). This is done by introducing a positive slack variable ξ_i, $i=1,2,\cdots L$:

(1.1) $$x_i \cdot w + b \geq +1 - \xi_i \quad (y_i = +1)$$

(1.2) $$x_i \cdot w + b \leq -1 + \xi_i \quad (y_i = -1)$$

Which can be combined into

$$y_i(x_i \cdot w + b) - 1 + \xi_i \geq 0$$

where $\xi_i \geq 0$

In this soft margin SVM, data points on the incorrect side of the margin boundary have a penalty that increases with the distance from it. As we are trying to reduce the number of misclassifications, a sensible way to adapt our objective function from previously, is to find:

$$\min \frac{1}{2}\|w^2\| + C\sum_{i=1}^{L}\xi_i$$
$$s.t. \quad y_i(x_i \cdot w + b) - 1 + \xi_i \geq 0$$

where the parameter C controls the trade-off between the slack variable penalty and the size of the margin.

When applying SVM to nonlinearly dataset, we need define a feature mapping function $x \mapsto \phi(x)$.

There are three common kernel functions:

○,1Polynomial Kernel

$$k(x_i, x_j) = (x_i \cdot x_j + a)^b$$

○,2Radial Basis Kernel

$$k(x_i, x_j) = e^{-\left(\frac{\|x_i - x_j\|^2}{2\sigma^2}\right)}$$

○,3Sigmoidal Kernel

$$k(x_i, x_j) = \tanh(ax_i \cdot x_j - b)$$

where, a and b are two parameters.

Genetic algorithm[7]

Genetic algorithm is presented by Prof. Holland in 1975. The algorithm has been widely implemented in a computer simulation in which a population of abstract representations (called chromosomes or the genotype of the genome) of candidate solutions (called individuals, creatures, or phenotypes) to an optimization problem evolves toward better solutions. Traditionally, solutions are represented in binary as strings of 0s and 1s, but other encodings are also possible. The evolution usually starts from a population of randomly generated individuals and happens in generations. In each generation, the fitness of every individual in the population is evaluated, multiple individuals are stochastically selected from the current population (based on their fitness), and modified (recombined and possibly randomly mutated) to form a new population. The new population is then used in the next iteration of the algorithm. Commonly, the algorithm terminates when either a maximum number of generations has been produced, or a satisfactory fitness level has been reached for the population. If the algorithm has terminated due to a maximum number of generations, a satisfactory solution may or may not have been reached.

Genetic algorithms find application in bioinformatics, phylogenetics, computational science, engineering, economics, chemistry, manufacturing, mathematics, physics and other fields.

Text Classification[8]

Text classification (also known as text categorization or topic spotting) is the task of automatically sorting a set of documents into categories from a predefined set. This task has several applications, including automated indexing of scientific articles according to predefined thesauri of technical terms, filing patents into patent directories, selective dissemination of information to information consumers, automated population of hierarchical catalogues of Web resources, spam filtering, identification of document genre, authorship attribution, survey coding, and even automated essay grading. Automated text classification is attractive because it frees organizations from the need of manually organizing document bases, which can be too expensive, or simply not feasible given the time constraints of the application or the number of documents involved. The accuracy of modern text classification systems rivals that of trained human professionals, thanks to a combination of information retrieval (IR) technology and machine learning (ML) technology.

Vector space model[9]

Vector space model (or term vector model) is an algebraic model for representing text documents (and any objects, in general) as vectors of identifiers, such as, for example, index terms. It is used in information filtering, information retrieval, indexing and relevancy rankings. Its first use was in the SMART Information Retrieval System.

A document is represented as a vector. Each dimension corresponds to a separate term. If a term occurs in the document, its value in the vector is non-zero. Several different ways of computing these values, also known as (term) weights, have been developed. One of the best known schemes is tf-idf weighting (tf_d is term frequency of term *t* in document *d*, idf is named inverse document frequency). The definition of term depends on the application. Typically terms are single words, keywords, or longer phrases. If the words are chosen to be the terms, the dimensionality of the vector is the number of words in the vocabulary, which is the number of distinct words occurring in the corpus.

3. GA-SVM algorithm

Algorithm introduction

There are two methods that are used to combine Genetic algorithm and SVM.

1) Dealing with dataset[10][11]

The initial training dataset are optimized with GA in order to find a sample subset including the important samples that can preserve or improve the discrimination ability of SVM. Training on the subset is equal to that on the initial sample sets. The training time is greatly shortened.

There are two available sets of training data: class1 $\{z_1, z_2, \cdots z_n\}$ and class2 $\{z'_1, z'_2, \cdots z'_n\}$. z_i or z'_j is an example, or one feature vector for SVM. An intuition idea is to find out the important examples that affect the classification results greatly. If these feature vectors are removed, the separating boundary changes the most. The key important question is how to find out these important training data from all the examples with GA.

2) Defining parameters[12-14]

The value of parameters in Support Vector Machines is important to algorithm's performance. Ángel Kuri-Morales and Iván Mejía-Guevara presented a methodology to train SVM where the regularization parameter (C) was determined automatically via an efficient Genetic Algorithm in order to solve multiple category classification problems.

In previous works, the support vectors have been determined from the application of Lagrange Multipliers, but are not applicable to search for "C". In fact, GA are used to solve the constrained QP. One advantage of using GA for this kind of problems is that restrictions are not imposed in the form of the objective function: a neither the objective function nor the constraints of the problem must be derivable in order to solve problems. In some cases, each individual represents a LM ($\alpha_i, i = 1,2,3,\cdots N$), where N is the number of points in the training set for the dual SVM problem.

This algorithm combing GA and SVM has been applied many fields, such as fault detection [15], protein sequences classification [16][17], network intrusion detection[18], daily flow forecasting [19], Short-term load forecasting[20], Evaluation of competitiveness of power plants[21], stock index forecasting[22].

The combing algorithm has better performance in classification and forecasting. This paper presents a new algorithm-EGA-SVM that combine EGA and SVM. An issue considered in the performance is that traditional GA itself has the problem: it is possible that some better solution found in previous steps may be lost because of the genetic operation. For reserving these better solutions, the algorithm should have the memory ability. Elite survival strategy is employed in combing algorithm, EGA-SVM.

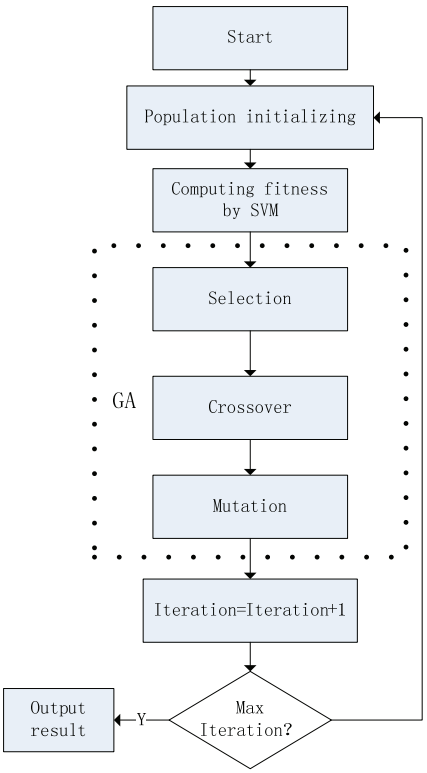

Fig.1 The flowchart of GA-SVM

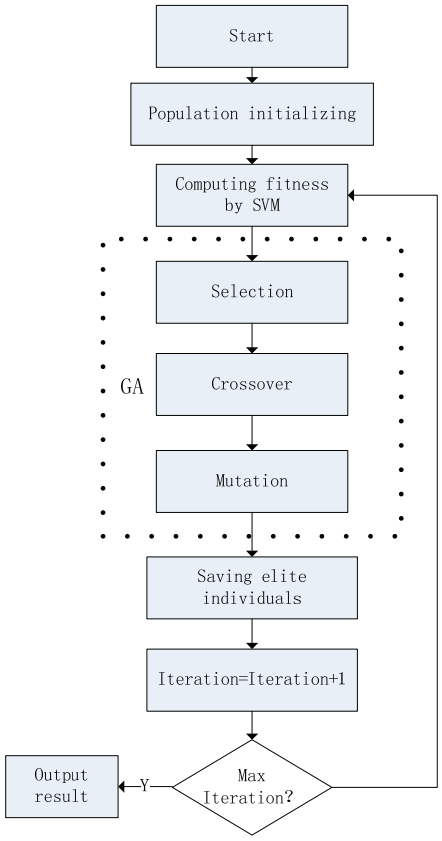

Fig. 2. The flowchart of EGA-SVM

Algorithm Description

EGA-SVM algorithm can be showed as following,

Step1 Create initial population randomly, the number of population is popsize

Step2 Classify dataset by SVM, calculate the Accuracy Rate that will be the value of fitness in Genetic algorithm

Step3 Genetic operation: selection, crossover and mutation

Saving elite individuals

Step4 If not the maximal iteration N, then to Step2;

Or, quit algorithm and output the classification result.

The flowchart of GA-SVM algorithm and EGA-SVM are showed in figure1 and figure 2.

4. Simulation Experiment and Discussion

Datasets for numerical experiments

Dataset 1: Iris dataset

The iris dataset consists of 150 data points with four attributes, and it is stored in a text file, which can be downloaded from the university of california irvine machine learning repository, http://archive.ics.uci.edu/ml/datasets/Iris. It is one of the best known datasets to be found in the pattern recognition literatures for assessing classification ability of different algorithms. Figure 3 shows the distribution of Iris data points in first three dimensions. This paper chooses forty samples as train dataset from each cluster and the other as test dataset.

Dataset 2: Text dataset

For validating the validity of the new algorithm, EGA-SVM, one hundred pieces of web pages in Chinese news are chosen as text dataset. At first, the dataset is classified by manual, including five clusters, related to finance, sport, military, science and culture. There are twenty pieces of news in each class. Five pieces of news are chosen for each class, so twenty-five pieces of texts are used as testing dataset. The rest of text dataset, seventy-five pieces of texts, are training dataset.

Before applying the new algorithm, text dataset need be transferred into text-terms matrix by vector space model. The size of the text-terms matrix that is transferred by the text dataset is 100*14924, that is one hundred pieces of text and 14924 terms.

Experiment result

For the compare of classifying performance among traditional SVM, GA-SVM and EGA-SVM, three algorithms are run several times respectively. In iris dataset, three algorithms are run ten times respectively. Accuracy rate of classification and running times are recorded. In text dataset, they are run a dozen times and first five best results, accuracy rate of classification, are chosen for algorithm's evaluation.

The programs of three algorithms are written based on Matlab 7.6 (version R2008a) and run on a computer with 2.0 GHz CPU, 2GB DDR RAM. The experiment results are showed in table 1, table 2 and figure 4.

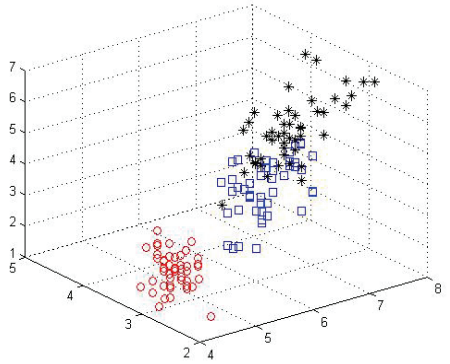

Fig. 3. The distribution of Iris data (three dimensions)

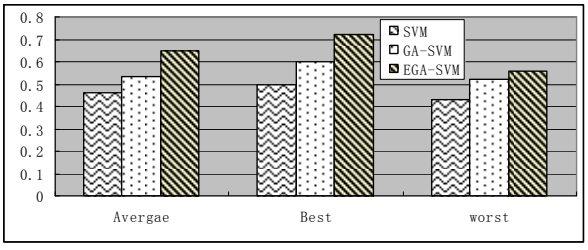

Fig. 4. The Compare of accuracy rate of classification among three algorithms in Text dataset

Table1 the Compare of Results among three algorithm in Iris dataset

	Accuracy Rate of classification			Running time		
	Best	Worst	Avg.	Best	Worst	Avg.
SVM	1	0.833	0.9033	0.1406	0.2813	0.19064
GA-SVM	1	1	1	6.9063	7.8906	7.3172
EGA-SVM	1	1	1	6.6094	7.3594	6.93439

Table2 the Compare of Results among three algorithm in Text dataset

	1	2	3	4	5	Best	Worst	Average
SVM	0.43	0.5	0.47	0.47	0.43	0.50	0.43	0.460
GA-SVM	0.52	0.60	0.52	0.52	0.52	0.60	0.52	0.536
EGA-SVM	0.64	0.64	0.72	0.68	0.56	0.72	0.56	0.648

Discussion

In iris dataset, GA-SVM and EGA-SVM can get better classification performance. Though traditional SVM need lowest time to run, there is worst in accuracy rate of classification among three algorithms.

In text dataset, which consists of one hundred pieces of web pages, the maximal accuracy rate of classification of traditional SVM is 50% in a dozen times, the value of GA-SVM is 60%, yet, the value of EGA-SVM is 72%. And the average accuracy rate of traditional SVM is 46%, yet, the value of GA-SVM is 53.6%, and the value of EGA-SVM is 64.8%. So the EGA-SVM is better than traditional SVM and GA-SVM in classification performance.

5. Conclusions and future work

This paper presents a new algorithm combing elite genetic algorithm and support vectors machine. It uses good optimization performance of support vector machines to improve classification performance of genetic algorithm with elite strategy. Iris dataset and a text dataset are chosen to validity performance of the combing algorithm. It's obviously that the hybrid algorithm can be applied to classify literatures in the field of electrical engineering. Future study direction will focus on the effect to performance when related parameters, such as crossing-over rate, mutation rate, size of population, etc., have different values, and improve computational efficiency of the new algorithm further.

Acknowledgment

The authors would like to thank anonymous reviewers for their constructive and enlightening comments, which improved this manuscript.

This work has been supported by grants from Program for Excellent and Creative Young Talents in Universities of Guangdong Province (No. LYM10097).

REFERENCES

[1] ZHONG Jiang, WEN Luo-Sheng, FENG Yong, YE Chun-Xiao and LI Zhi-Gu. Study on the Web Classification Based on Proximal Support Vector Machine. Computer Science, 2008,35(3), pp.167-169,202.

[2] LI Tao， W ANG Jun-pu and XU Yang. A Rough Set Method for Web Classification. Mini-micro System, 2003,24(3),pp.520-522.

[3] FENG He-long and XIA Sheng-ping. Web Page Classification Method Based On RSOM-Bayes. Computer Engineering, 2008,34(13),pp.61-63.

[4] Liu Li-zhen, HE Hai-jun, Lu Yu-chang and Song Han-tao. Application Research of Support Vector Machine in Web Information Classification. MINI-MICRO SYSTEM, 2007,28(2), pp.337-340.

[5] Tristan Fletcher. Support Vector Machines Explained. http://www.tristanfletcher.co.uk/SVM%20Explained.pdf. [2009-10-6]

[6] Support vector machine. http://en.wikipedia.org/wiki/Support_vector_machine [2009-10-6]

[7] Genetic algorithm. http://en.wikipedia.org/wiki/Genetic_algorithm. [2009-10-6]

[8] Text classification. http://en.wikipedia.org/wiki/Text_classification[2009-10-6]

[9] Vector space model. http://en.wikipedia.org/wiki/Vector_space_model [2009-10-6]

[10]Zhang, X.R., and Liu, F.: 'A patten classification method based on GA and SVM', 2002 6th International Conference on Signal Processing Proceedings, Vols I and Ii, 2002, pp. 110-113

[11] Liu, J.J., Cutler, G., Li, W.X., Pan, Z., Peng, S.H., Hoey, T., Chen, L.B., and Ling, X.F.B.: 'Multiclass cancer classification and biomarker discovery using GA-based algorithms', Bioinformatics, 2005, 21, (11), pp. 2691-2697

[12] Liu, S., Jia, C.Y., and Ma, H.: 'A new weighted support vector machine with GA-based parameter selection', Proceedings of 2005 International Conference on Machine Learning and Cybernetics, Vols 1-9, 2005, pp. 4351-4355

[13] ZHANG X R, LIU F. A patten classification method based on GA and SVM. 6th International Conference on Signal Processing Proceedings, Vols I and Ii, 2002, pp.110-113.

[14] KURI-MORALES A, MEJIA-GUEVARA I. Evolutionary training of SVM for multiple category classification problems with self-adaptive parameters. Advances in Artificial Intelligence - Iberamia-Sbia 2006, pp.(329-338).

[15] Nguyen, N.T., and Lee, H.H.: 'An Application of Support Vector Machines for Induction Motor Fault Diagnosis with Using Genetic Algorithm', Advanced Intelligent Computing Theories and Applications, Proceedings, 2008, 5227, pp. 190-200

[16] Zhao, X.M., Huang, D.S., Cheung, Y.M., Wang, H.Q., and Huang, X.: 'A novel hybrid GA/SVM system for protein sequences classification', Intelligent Data Engineering and Automated Learning Ideal 2004, Proceedings, 2004, 3177, pp. 11-16

[17] Li, S.T., Wu, X.X., and Hu, X.Y.: 'Gene selection using genetic algorithm and support vectors machines', Soft Computing, 2008, 12, (7), pp. 693-698

[18] Kim, D.S., and Park, G.S.: 'Modeling network intrusion detection system using feature selection and parameters optimization', Ieice Transactions on Information and Systems, 2008, E91D, (4), pp. 1050-1057

[19] Jianzhong, W., Ling, L., and Juan, C.: 'Combination of genetic algorithm and support vector machine for daily flow forecasting', 2008 Fourth International Conference on Natural Computation (ICNC), 2008, pp. 31-35

[20] Ma, L.H., Zhou, S.G., and Lin, M.: 'Support Vector Machine Optimized with Genetic Algorithm for Short-term Load Forecasting', Kam: 2008 International Symposium on Knowledge Acquisition and Modeling, Proceedings, 2008, pp. 654-657

[21] Wei, S., and Jie, Z.: 'Evaluation of competitiveness of power plants based on optimized SVM using GA and AIS', 2008 International Conference on Risk Management & Engineering Management, 2008, pp. 648-652

[22] Huang, S.C., and Wu, T.K 'Integrating GA-based time-scale feature extractions with SVMs for stock index forecasting', Expert Systems with Applications, 2008, 35, (4), pp. 2080-2088

***Authors**: Dr. Xiaoyong LIU, Department of Computer Science, Guangdong Polytechnic Normal University, Guangzhou, Guangdong, 510665, China, E-mail: lxyong420@126.com; Mrs. Hui FU, Department of Computer Science, Guangdong Polytechnic Normal University, Guangzhou, Guangdong, 510665, China, E-mail: lindafh819@126.com.*