Applying a Novel Scheme in Risk Assessment for Port State Control Inspection

Zhong Gao[1], Huike Li[1], Guanming Lu[1] and Daquan Gu[2]
[1]*Collegel of Telecommunications and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing, China*
gaozhong527@gmail.com,
[2]*College of Meteorology, PLA University of Science and Technology, Nanjing, China*

Abstract

Port State Control (PSC) inspection is the most important mechanism to ensure world marine safe. Recently, some SVM-based risk assessment schemes have been presented in the world. They estimate the risk of each candidate ship based on its generic factors and history inspection factors to select high-risk one before conducting on-board PSC inspection. However, how to improve the performance of the PSC inspection under the situation of noisy data when applying SVM is still a challenging problem. In this paper, we propose a new approach for PSC inspection, which uses a novel Support Vector Machine and K-nearest neighbor (KNN-SVM) to remove noisy training examples and Bag of Words (BW) to extract some new target factors for the PSC inspection database. The experimental results show that the generalization performance and the accuracy of risk assessment are improved significantly compared to that of the traditional SVM classifier, and adapt to engineering applications.

Keywords: Port State Control, Support Vector Machine, K-nearest neighbor, Bag of Words.

1. Introduction

Port State Control (PSC) is established to ensure that foreign ships are seaworthy, do not pose a pollution risk, provide a healthy and safe working environment and comply with relevant international conventions [1]. Various regions have established the organization of regional cooperation on PSC such as Paris Memorandum of Understanding on PSC (Paris MOU) and Tokyo Memorandum of Understanding on PSC (Tokyo MOU). The implementation of PSC control in the past years effectively reduces marine risk in the past decade.

At present, there are some studies on intelligent risk assessment for improving PSC inspection. These schemes make risk assessment based on the features recorded in PSC inspection database and their performance is better than Paris MOU PSC mechanism [1]. However, the fixed weighting and scoring parameters adopted in these schemes are manually adjusted based on the expert knowledge. The performance of these schemes is difficult to improve.

In this paper, we present a risk assessment scheme, which is based on the study on the initial target factors suggested by Paris MOU and Tokyo MOU. The initial set is then expanded by adding more target factors extracted by Bag of Words from the PSC inspection web pages. These target factors are used to build a risk assessment component based on Support Vector Machine which has high generalization ability. However, SVM is very sensitive to noisy training data, which can degrade its classification accuracy. So we put forward a noise-tolerant algorithm to construct classifiers for risk assessment, which combines the current two top performing methods: support vector machines and K-nearest neighbor method. Using inspection records as the training data, the scheme is applied to estimate the risk of candidate ships and to select high-risk ones. The ships with high risk and medium-high risk are provided to the port officers to help them selecting ships with potential marine risks with good accuracy. Evaluations show that this scheme outperforms the risk estimation mechanisms adopted in Paris MOU and Tokyo MOU.

2. Related works

The Paris MOU inspection records are given in http://www.parismou.org/ParisMOU/Inspection+Database/xp/menu.3973/default.aspx. It maintains about 83,000 inspection records conducted by the Paris MOU ports from Jan. 2003 to Jan. 2009. The Tokyo

MOU inspections are provided in http://www.tokyo-mou.org/. We must firstly extract specific data from those web pages. After specifying the features of data by profiles, the relevant web content can be extracted from a batch of text documents which are presented in the same format. Then we select the target factors from the extracted text features and build up a PSC inspection database.

2.1. Web page text representation

The first of this pre-work is to extract a group of suitable feature terms from HTML pages. To accomplish automatic text representation, the set of documents, typically strings of characters, has firstly to be converted to an acceptable representation that the learning machine can handle. The most common, simple and successful representation so far is the vector space model, also known as the Bag of Words. Each document is indexed with the bag of the terms occurring in it, i.e., a vector with one component for each term occurring in the whole collection, having as value the number of times the term occurred in the document. Each document is thus represented as a point in a vector space with one dimension for every term in the vocabulary, thus loosing word order information. The number of times a term occurs in a document is usually referred to as term frequency or tf . The method described so far gives equal importance to all terms. It can be however more effective to weight terms according to their discriminative power within the document collection. This is usually done based on the idea that terms occurring in fewer documents are better selectors. The document frequency $df(t)$ of a term t is the number of documents in the collection in which the term occurs. The inverse document frequency or $idf(t)$ is:

$$idf(t) = \log\left(\frac{|D|}{df(t)}\right) \quad (1)$$

where $|D|$ is the number of documents in the collection.

Vector components are weighted according to the $idf(t)$ of the corresponding term. Usually some monotonous function of the tf , such as the log or the square root, is used instead of the tf itself, to avoid giving more importance than appropriate to multiple occurrence of terms.

2.2. Target factor selection

The target factors considering in PSC inspection are classified into three types: generic (static) factor, dynamic factor and history factor. Generic factors include ship's name, code, date and place of building, and type and design. This information does not change during the life of a ship. Furthermore, the information relating to the management of the ships, including their registers, managers, insurers, crew, maintenance and so on are also considered as generic factors. Dynamic factors refer to the changes of captain and sailors on the ship and the types of cargo loaded in each shipping. History factor is the inspection history of a ship including the place and date of inspections, the deficiencies detected and detention information. In this section, we investigate and select corresponding 15 target factors in Pairs MOU and Tokyo MOU including some new target factors based on the statistical analysis on text data extracted from Paris MOU and Tokyo MOU PSC inspection records web pages.

3. Risk assessment based on the proposed algorithm

Support vector machine and k-nearest neighbor are combined to form a KNN-SVM to deal with the noisy training data and enhance the performance of PSC inspection

3.1. SVM and problem with it

Support Vector Machine (SVM) is a powerful classification algorithm [2]. Different from the classical Hidden Markov Model and Maximum Entropy Modeling that require careful feature selection as they do not provide methods to automatically select optimum features from a given feature set, classification based on SVM has the capacity to identify optimum features through training procedure. In the basic framework of SVM, SVM identifies an optimal hyper plane to separate training data into two classes, positive cases and negative cases. SVM takes the strategy which maximizes the margin between critical samples and the separating hyper plane. In particular, SVM achieves high generalization even with training data of a very high dimension. In practice, even in the case that training data can not be linearly separated because of the existence of some noise, the separating linear hyper plane still can be built by allowing some misclassifications. Furthermore, by introducing the Kernel function, SVM handle non-linear feature spaces, and carry out the training considering combinations of more than one feature.

SVM is adopted to build a risk assessment component by integrate the 15 target features into a linear expression [3]. The 15 vectors corresponding to the proposed 15 target factors for each candidate ship in the training data are calculated. The regression function in SVM is trained by values of these features. This SVM function is then applied to the testing data. As such, it determines whether a ship should be inspected since it has a high risk.

Although support vector machine has shown promising performance in many applications, the process of learning a SVM classifier is easily affected by noisy training data. Noisy data points usually locate close to the hyper-plane. They often become support vectors easily during the learning process. And only the support vectors determine the position of hyper-plane. So if there is noisy data in training set, it will change the position of the optimal hyper-plane and degrade the ability of generalization of SVM classifiers, and finally bring down the classification accuracy of SVM classifiers.

3.2. K-Nearest Neighbor

K-nearest neighbor (KNN) method is a relatively mature machine learning theory first proposed by Cover and Hart in 1968. A KNN classifier determines the class label of test example based on its K neighbors that are closet to it. Any test example is classified into the class which has the maximize number of examples among its K closet neighbors. Compared to support vector machine, a KNN classifier still has high classification accuracy if there is noisy data in the training set. That also means that KNN can deal with noisy data better. But it's time-consuming during the process of classification. However, K-nearest neighbor together with support vector machines is still the two top performing machine-learning methods for text classification. Thus, this paper employs KNN method to improve the capability of SVM classifiers.

3.3. The combined classifier

Since a SVM classifier is sensitive to noisy training data, these noisy training data should all be discarded before learning a SVM classifier at best. In this paper, we employ K-nearest neighbor method to edit the training data for our experiments; and term this noise-tolerant support vector machine as KNN-SVM algorithm. KNN-SVM suggests dividing the training examples into s splits in which each class contains the average number of training examples, and constructing a KNN classifier for each split S_i. Then it classifies all the examples in S, using the next KNN classifier $f(i+1)$, and discards all the misclassified training examples. Finally, it uses all the remaining examples to build a SVM classifier.

For this KNN-SVM, the computational complexity mainly contains two parts: editing the noisy training data using KNN and training the final SVMs. The computational complexity for editing the noisy training data is $O(\mathrm{m} N^2)$ where m is the size of its dimension and N is the number of training samples. The computational complexity for training the final SVM s is $O(N^3)$. So the computational complexity for this algorithm is $O(\mathrm{m} N^2 + N^3)$, where N is the number of the training samples.

4. Experimental results

4,000 ships with more than 4 inspection records are selected from Paris MOU PSC database, respectively. These data are further divided into 2 parts in which inspection records of 3,000 ships are used as training data, and the rests are used as testing data. We use the last inspection as the verification of the decision and the previous inspection result as the history data. If a ship is decided as with high risk and the inspection leads to a detention, it is regarded as a successful risk assessment. Therefore, the accuracy of detention forecast is used as the evaluation criteria in the experiment. In the following section, the training data for the Paris and Tokyo MOU PSC inspections are labeled as PM_Train, and TM_Train, respectively. Similarly, the testing data for the two groups of data are labeled as PM_Test and TM_Test. Note that the cross validation is considered in the experiments. The achieved average accuracy is given the following experiments.

The performance of the proposed KNN-SVM is investigated in the MATLAB environment. In the proposed scheme, we presented a non-linear SVM with a Gaussian kernel. In the first experiment, we estimate the performance of suggested risk assessment of Paris MOU and Tokyo MOU PSC, respectively. They are regarded as the baseline. The achieved risk assessment accuracies are given in Table 1.

Table 1. Risk assessment accuracy of Pairs MOU PSC and Tokyo MOU PSC

	Paris MOU PSC		Tokyo MOU PSC	
	SVM	KNN-SVM	SVM	KNN-SVM
PM_Test	7.21%	13.97%	9.12%	16.35%
TM_Test	8.24%	15.63%	8.96%	15.84%

From the table, we can clearly see that when applying KNN-SVM the classification accuracy is increasing apparently. Moreover, it is observed that the assessment accuracy achieved by Tokyo MOU PSC specifications is higher than the one achieved by Paris MOU PSC specifications.

In the second experiment, we evaluate the SVM-based risk assessment system with all of the 15 factors, i.e., the complete system. The achieved risk assessment accuracies are given in Table 2.

Table 2. Risk assessment accuracy of the final scheme

Training Data	Testing Data	Accuracy	
		SVM	KNN-SVM
PM_Train	PM_Test	10.43%	17.39%
PM_Train	TM_Test	10.02%	18.52%
TM_Train	PM_Test	9.96%	15.60%
TM_Train	TM_Test	12.70%	18.97%
PM_Train+TM_Train	PM_Test	13.84%	20.52%
PM_Train+TM_Train	TM_Test	14.44%	20.93%

This result indicates that these factors are effective in PSC risk assessment and KNN-SVM has the better classification accuracy than that of traditional SVM. The final achieved assessment accuracy for PM_test and TM_Test are 20.52% and 20.93%, respectively. It is significantly improvement from the scheme, which is clearly an acceptable result for us.

5. Conclusion

Port State Control is playing an increasingly important role in world marine safety. Therefore, in this paper, we proposed some new target factors and investigated a KNN-SVM classifier for estimating risks of candidate ships in PSC inspection. The structure of KNN-SVM is simple and it has good generalization ability. These target factors are incorporated with the factors adopted in Paris MOU and Tokyo MOU to develop a risk assessment scheme for helping the port officers selecting the highest risk ships. Evaluations show that the proposed system obviously improves the accuracy of risk assessment compared to that of the traditional SVM classifier. It can be concluded that in this experiment and possibly in many other cases, the target factors that we extract are considered to be more effective for the KNN-SVM classifier and this combined classifier based on these features could lead to a more satisfied results in PSC inspection.

6. Acknowledgement

The authors would like to thank the research staffs from the Computing Center of PLA University of Science and Technology for their kindly and powerful help.

7. References

[1] IMO News 2000 No.1 Port State Control: Verify Safety Standards Worldwide, 2000.

[2] Cortes C. and Vapnik V. N. “Support Vector Networks”, *Machine Learning*, vol. 20, pp. 273-297, 1995.

[3] B. Scholkopf, et al. “Support Vector Regression with Automatic Accuracy Control”, *Proceeding of 8th International Conference on Artificial Neural Networks*, pp.111-116, 1998.

[4] Soares C.G. and Teixeira A.P., “Risk assessment in maritime transportation”, *Reliability Engineering and System Safety*, vol.74, pp. 299-309, 2001.

[5] Shen Q.X. and Chen W.J, “The Design and Implementation of Automatic Ship Selection in Port State Control”, *Shipping Management*, Vol.2003, no.2, pp.32-34, 2003.

[6] Yang Z. *The Quantify Research on the Ship Selection in Port State Control Inspection*, Master Thesis, Shanghai Maritime University, 2004.

[7] Gramling, R., Forsyth, C.J. and Wooddell, G., “Expert informants and relative risk: A methodology for modeling waterways”, *Risk Analysis*, Vol. 18, No.5, pp. 557-562, 1998

[8] Rosqvist, T. and Tuominen, R., 4, Qualification of formal safety assessment: an exploratory study, *Safety Science*, Vol. 42, No.2, pp. 99-120, 2004

[9] T. Joachims. A Statistical Learning Model of Text Classification with Support Vector Machines. *Proc. of SIGIR-01*, New Orleans. ACM Press, New York, 2001

[10] Sii, H.S., Ruxton, T. and Wang, J., A fuzzy-logic-based approach to qualitative safety modeling for marine systems, *Reliability Engineering and System Safety*, Vol. 73, No. 1, pp. 19-34, 2001