The Second International Symposium on Optimization and Systems Biology (OSB'08)
Lijiang, China, October 31– November 3, 2008

Fault Diagnosis by Using Selective Ensemble Learning Based on Mutual Information

Tian-Yu Liu[1] Guo-Zheng Li[2]

[1]School of Electric, Shanghai Dianji University, Shanghai 200040, China
[2]Department of Control Science and Engineering, Tongji University, Shanghai, 201804 China

Abstract Fault diagnosis on diesel engine is a difficult problem due to the complex structure of the engines and the presence of multi-excite sources. There have been previous attempts to solve this problem by using artificial neural networks and others methods. In this paper, a novel algorithm named MISEN (Mutual Information based Selective Ensemble) is proposed to improve diagnosis accuracy and efficiency. MISEN is compared with the general case of bagging and GASEN, a baseline method, namely Genetic Algorithm Based Selective ENsemble, on UCI data sets. Then, MISEN is used to diagnose the diesel engine. Computational results show that MISEN obtains higher accuracy than other several methods like bagging of neural networks and GASEN.

Keywords Fault diagnosis; Mutual information; Selective ensemble learning; Bagging; Support vector machines

1 Introduction

Diesel engines are a kind of power machine widely used in many fields. As one of the most widely used power sources, the diesel engine has been drawing more attention in the aspect of its fault diagnosis [1]. The diesel engine is the representative of back and forth machine, it is difficult to diagnose its faults by using traditional methods because of its complexity and multiplicity. There have been previous attempts to solve this problem by using artificial neural networks and others methods [2, 3]. But, there is a need to have a method that can diagnose more than one category of faults in a generic manner with high accuracy. In this paper, the relationship between the category of faults in the Diesel engine and vibration signals will be studied by a novel selective ensemble learning algorithm.

Ensemble learning is a learning paradigm where multiple component learners are trained for a same task by a same learning algorithm, and the predictions of the component learners are combined for dealing with future instances [5, 6]. Since an ensemble is often more accurate than its component learners, such a paradigm has become a hot topic in recent years.

In general, an ensemble is built in two steps, that is, training multiple component learners and then combining their predictions. According to the styles of training the component learners, current ensemble algorithms can be roughly categorized into two classes. The representative of the first category is Bagging [7]. The representative of the second category is AdaBoost [8]. It is worth mentioning that after obtaining multiple

learners, most ensemble algorithms employ all of them to constitute an ensemble. Although such a scheme works well, recently Zhou et al. [9] showed it may be better to ensemble some instead of all of them. They proposed an algorithm named GASEN, i.e. Genetic Algorithm based Selective ENsemble, which trains several individual neural networks and then employs genetic algorithm to select an optimum subset of individual neural networks to constitute an ensemble. Experiments show the performance of GASEN is excellent by using different learning machines. But GASEN uses genetic algorithm as an individual selection method, whose computational complexity is $O(2^n)$, n is the number of individuals, therefore, GASEN need much computation capacity. In order to improve the efficiency of GASEN, at the same time to improve the generalization ability, here we propose MISEN, namely Mutual Information based Selective ENsemble, which uses mutual information as the individual selection method, since the computational complexity of mutual information is $O(n^2)$, it is much lower than that of genetic algorithm. Support vector machines are state-of-the-art learning machines and have been studied widely as base learning machines of bagging [10], so using them as base machines makes the conclusion more reliable.

The rest of this paper is organized as follows. In Section 2, the used diesel engine data set for fault diagnosis is briefly described. In Section 3, we introduce the mutual information based individual selection method and present our MISEN algorithm. In Section 4, experimental results on UCI data sets and the diesel engine data set using MISEN are compared with that of GASEN. At last, conclusions are given in Section 5.

2 A diesel engine data set for fault diagnosis

In this paper, the diesel engine data set is from the original vibration signals sampled from a 4135 engine surface. The rated engine power of 4135 diesel engine is 80 hp and the rated engine speed is 1500 rpm. Three sampling points are selected for collecting vibration signals. They are located at the first cylinder head, the second cylinder head and the center of the piston stroke, on the surface of the cylinder block. Then, a method is used to discretize the attributes extracted from vibration signals. The data set is composed of 18 attributes (six attributes for each sampling point) and four states (normal state; intake valve clearance is too small; intake valve clearance is too large; exhaust valve clearance is too large). Among these four states, three fault types are obtained by deliberately introducing the corresponding fault conditions into the intake valve and exhaust valve on the second cylinder head. The attributes present the information contained in vibration signals both from the frequency domain and time domain[11].

3 Computational Methods

3.1 Mutual information

Mutual information (MI) describes the statistical dependence of two random variables or the amount of information that one variable contains about the other and it is a widely used information theoretic measure for the stochastic dependency of discrete random variables. It has been used to select features for classification problems [12], i.e. to select a subset of variables to predict the class variable. It is alternatively referred to as relative entropy or trans-information. The mutual information between two variable M and N is defined in terms of the probability distribution of intensities as follows:

$$I(M:N) = \sum_{m\in M}\sum_{n\in N} p\{m,n\} \lg \frac{p\{m,n\}}{p\{m\}p\{n\}}. \quad (1)$$

where $p\{m,n\}$ is the combined probability distribution of intensities of two variables M and N. $p\{m\}$ and $p\{n\}$ respectively are the individual probability distribution of intensities of M and N. The computational complexity of mutual information algorithm is $O(n^2)$, n is the length of R,S.

3.2 MISEN

To reduce the computational time, we use mutual information instead of genetic algorithm as the individual selection method, and propose a mutual information based selective ensemble algorithm, named MISEN [4].

The procedure of MISEN is shown in Algorithm 1. Briefly, MISEN first employs bootstrap to generate many models, then, these models are ranked by using the mutual information values calculated from the model outputs on the validation data set and the target vector of the validation data set. The best percent P models are selected to generate the ensemble model instead of using all the models in bagging, where P is a proportion pre-set by hand.The final classification decision is based on the majority voting of the ensemble model.

Algorithm 1 The MISEN algorithm

Input: Training data set S, validation set V, learner L, population size T, and proportion P
Output: Ensemble model N^*

```
Begin
for i = 1 to T do
    Generate a subset S_i by bootstrap
    Train the learner L to generate a model N_i on S_i
end for
for i = 1 to T do
    Test N_i on V and obtain Output(N_i)
    Compute the mutual information value M_i between
    Output(N_i) and the target vector in V by using Eq. (1)
end for
Rank N according to M in descending order
Select the first INT(P*T) models to generate the final ensemble model N*
End
```

4 Experiments on UCI data sets

4.1 The used UCI data sets

MISEN is compared with GASEN and Bagging on eleven data sets selected from UCI machine learning repository [13]. These data sets have been extensively used in testing the performance of diverse kinds of learning systems. To make them suitable for our algorithms, features and instances with missing values are removed and the nominal values are changed to be numerical in all data sets. Then, all the features are transformed into the interval of [-1, 1] by an affine function. The information of the UCI data sets used in our experiments are listed in Table 1.

Table 1: The properties of the UCI data sets for comparison

Data set	classes	features	instances
backup	19	35	683
audio	24	70	226
processed-Cleveland	5	13	303
processed-Hungarian	2	13	294
soybean-large	19	34	307
statlog-heart	2	13	270
glass	6	9	214
voting-records	2	16	435
Ionosphere	2	34	351
breast-cancer-Wisconsin	2	9	699
tic-tac-toe	2	9	958

4.2 Experimental Results

The hold out method is used to validate the results. Experiments are repeated fifty times on each data set. According to the advices of Valentini and Dietterich [10], the same pair of parameters for support vector machines, $C = 100, \sigma = 10$, is used and the number T of individuals for bagging is 20. The proportion P of MISEN is 75%.

The statistical prediction accuracy obtained on all data sets using MISEN, GASEN and Bagging are shown in Table 2, from which we can see that the results of accuracy obtained by MISEN are slightly better than those by GASEN, and both results by MISEN and GASEN are better than those by bagging.

The computational time of MISEN and GASEN during the process of selecting individuals are shown in Table 3, from which we can see that the computational time of MISEN is rather less than that of GASEN in all eleven data sets, the ratios of the computational time between MISEN and GASEN range from 0.09 to 0.57. The average value of ratios is 0.31, this means that MISEN uses less one third of the computational time of GASEN for selecting individuals on all the eleven data sets.

Table 2: Statistical accuracy by MISEN, GASEN and Bagging(%)

Data set	MISEN	GASEN	Bagging
backup	92.07 ± 1.17	91.85 ± 1.43	90.48 ± 2.27
audio	75.84 ± 3.99	75.62 ± 4.01	75.78 ± 3.63
processed-Cleveland	53.68 ± 2.99	53.39 ± 3.16	50.15 ± 3.30
processed-Hungarian	78.04 ± 2.56	77.63 ± 2.69	75.70 ± 3.07
soybean-large	85.04 ± 3.30	84.71 ± 1.08	83.83 ± 3.12
statlog-heart	78.50 ± 3.15	78.41 ± 2.91	75.94 ± 3.55
glass	65.19 ± 4.07	64.74 ± 4.46	61.42 ± 4.38
voting-records	94.94 ± 1.19	94.71 ± 1.23	94.12 ± 1.24
Ionosphere	89.21 ± 2.71	88.92 ± 2.72	87.55 ± 2.96
breast-cancer-Wisconsin	94.18 ± 1.11	93.69 ± 1.52	91.11 ± 1.95
tic-tac-toe	97.66 ± 0.72	97.20 ± 0.89	97.40 ± 0.97
Average	82.21 ± 2.45	81.89 ± 2.37	80.31 ± 2.78

Table 3: Average computational time of MISEN and GASEN during the process of selecting individuals

Data set	MISEN(s)	GASEN(s)	ratio
backup	18.23 ± 1.94	33.71 ± 3.71	0.54
audio	1.99 ± 0.09	14.23 ± 0.44	0.14
processed-Cleveland	0.85 ± 0.07	4.98 ± 0.12	0.17
processed-Hungarian	0.31 ± 0.05	3.39 ± 0.09	0.09
soybean-large	3.04 ± 0.10	16.82 ± 0.70	0.18
statlog-heart	0.45 ± 0.16	5.08 ± 1.70	0.09
glass	0.87 ± 0.27	7.95 ± 2.68	0.11
voting-records	0.59 ± 0.05	3.88 ± 0.08	0.15
Ionosphere	0.61 ± 0.20	4.69 ± 1.31	0.13
breast-cancer-Wisconsin	1.69 ± 0.12	4.85 ± 0.08	0.35
tic-tac-toe	5.91 ± 0.66	10.30 ± 0.67	0.57
Average	3.14 ± 0.33	9.99 ± 1.05	0.31

5 Results on the diesel engine data set

The information of the diesel engine data set used in our experiments are listed in Table 4. The hold out method is used to validate the result. Experiments are repeated fifty times on this data set. The pair of parameters for support vector machines, $C = 100, \sigma = 10$, is used and the number T of individuals for bagging is 20. The proportion P of MISEN is 75%.

The statistical prediction accuracy obtained on this data set using MISEN, GASEN and Bagging are shown in Table 5, from which we can see that the results of accuracy obtained by MISEN are slightly better than those by GASEN, and both results by MISEN and GASEN are better than those by bagging. The computational time of MISEN and

Table 4: The properties of the diesel engine data set

Data set	classes	features	instances
diesel engine	4	18	37

Table 5: Statistical accuracy by MISEN, GASEN and Bagging(%)

Data set	MISEN	GASEN	Bagging
diesel engine	82.78 ± 8.48	81.22 ± 9.02	80.78 ± 8.69

GASEN during the process of selecting individuals are shown in Table 6, from which we can see that the computational time of MISEN is rather less than that of GASEN on diesel engine data set, the ratio of the computational time between MISEN and GASEN is 0.12. This means that MISEN uses one fifth or less of the computational time of GASEN for selecting individuals on the diesel engine data set.

Table 6: Average computational time of MISEN and GASEN during the process of selecting individuals

Data set	MISEN(s)	GASEN(s)	ratio
diesel engine	0.2481 ± 0.0222	2.1435 ± 0.0342	0.12

6 Conclusions

To improve the fault diagnosis accuracy and efficiency of the diesel engine, the selective ensemble learning algorithm of MISEN (Mutual Information based Selective ENsemble) is employed. MISEN further validates that selective ensemble learning only using some of the individuals obtains better generalization performance than ensemble methods using all of the individuals. The computational complexity of mutual information is lower than that of genetic algorithm. Experimental results show that MISEN using mutual information as the individual selection method consumes less computation time than GASEN using genetic algorithm. At the same time, MISEN obtains slightly better accuracy than GASEN does. In a word, MISEN may be an alternative method in the family of selective ensemble learning.

From the view of the fault diagnosis, the computational results imply that the relationship between the fault category and attributes can be modeled by neural networks based methods, of which MISEN is a choice. Further work need to apply MISEN to more diagnosis data sets.

Acknowledgements

This work was supported by Special Fund for Excellent Young Teacher Research Project of Shanghai Municipal Education Commission (No. sdj-07003), the National Nat-

ural Science Foundation of China (No. 20503015 and 60873129), Innovation Program of Shanghai Municipal Education Commission (No. 08YZ189), Leading Academic Discipline Project of Shanghai Municipal Education Commission (No. J51901), the STCSM "Innovation Action Plan" Project of China (No. 07DZ19726), the Shanghai Rising-Star Program (No. 08QA14032) and Systems Biology Research Foundation of Shanghai University.

References

[1] D.Huang,:Vibration analysis and fault diagnosis for diesel engine. Journal of Vibration Engineering 8(2) (1995) 144-149

[2] Zhang Xi-ning, Wen Guang-rui , Li Xin-yi:Vibration Feature Abstraction and Classification of Diesel Fault. International Journal of Plant Engineering and Management 8(1) (2003) 35-40

[3] Wang Cheng-dong , Wei Rui-xuan , Zhang You-yun, et al: Fault Diagnosis for a Diesel Valve Train Based on Time-Frequency Analysis and Probabilistic Neural Networks. International Journal of Plant Engineering and Management 9 (3) (2004) 155-163

[4] G.-Z. Li, J.Y. Yang. Chapter 6, Feature Selection for Ensemble Learning and Its Application, In: Machine Learning in Bioinformatics, New York, John Wiley & Sons, (2008) 135-155.

[5] Dietterich, T.: Machine-learning research: Four current directions. The AI Magazine 18 (1998) 97-136

[6] Schapire, R.: The strength of weak learn ability. Machine learning 5 (1990) 197-227

[7] Breiman, L.: Bagging predictors. machine learning. Machine learning 24 (1996) 123-140

[8] Bauer, E., Kohavi, R.: An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. Machine Learning 36 (1999) 105-139

[9] Zhou, Z.H., Wu, J.X., Tang, W.: Ensembling neural networks: many could be better than all. Artificial Intelligence 137 (2002) 239-263

[10] Valentini, G., Dietterich, T.: Bias-variance analysis of support vector machines for the development of svm-based ensemble methods. Journal of Machine Learning Research 5 (2004) 725-775

[11] Shen, L., Tay, F.E.H., Qu, L., Shen, Y.: Fault diagnosis using rough sets theory. Computers in Industry 43 (2000) 61-72

[12] Peng, H., Long, F., Ding, C.: Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transactions on Pattern Analysis and Machine Intelligence 27(8) (2005) 1226-1238

[13] Blake, C., Keogh, E., Merz, C.J.: UCI repository of machine learning databases. Technical report, Department of Information and Computer Science, University of California, Irvine, CA (1998) http://www.ics.uci.edu/ mlearn/MLRepository.htm.