Available online www.jocpr.com

Journal of Chemical and Pharmaceutical Research, 2015, 7(3):1287-1292

Research Article

ISSN : 0975-7384
CODEN(USA) : JCPRC5

Imbalance learning for fault diagnosis gearbox in wind turbine

Liu Tianyu

School of Electric, Shanghai Dianji University, Shanghai, China

---

ABSTRACT

Defect is one of the important factors resulting in gearbox of wind turbine, so it is significant to study the technology of defect diagnosis for gearbox. Class imbalance problem is encountered in the fault diagnosis, which causes seriously negative effect on the performance of classifiers that assume a balanced distribution of classes. Though it is critical, few previous works paid attention to this class imbalance problem in the fault diagnosis of gearbox. In imbalanced problems, some features are redundant and even irrelevant. These features will hurt the generalization performance of learning machines. Here we propose PSO (Particle Swarm Optimization based feature selection for Easy Ensemble) to solve the class imbalanced problem in the fault diagnosis of gear. Experimental results on UCI data sets and gearbox data set show that PSOEE improves the classification performance and prediction ability on the imbalanced dataset.

Key words: Particle Swarm Optimization; wind turbine gearbox; fault diagnosis; imbalanced data; ensemble learning

---

INTRODUCTION

Wind turbine running long in harsh natural environment outdoor, the failure rate is higher than that of conventional generators. According to incomplete statistics, at present average availability of wind turbine in China is generally lower than 95%[1], in addition to wind power access system does not have the conditions, high failure rate of wind turbine is one of the main factors, these factors led to the wind turbine maintenance costs become the main operation cost of wind farms, according to the wind power unit 20 years during the whole life cycle cost average calculation, wind turbine maintenance costs about 1.2 €/kWh[2], therefore reduce maintenance cost is an important way to improve the operating efficiency of the wind farm.

The gear box is the key components of wind turbine. The high failure rate of gear box is in each of the main components of the wind turbine. According to the statistics of failure [3] on the main components of the wind turbine of the UK renewable energy center, the gear box is the highest rate of components; failure rate has more than 60%.[4] Because of the wind turbine installed in the tens of meters high tower, repairing gear box is very inconvenient. Therefore, for reducing the maintenance cost, to strengthen the monitoring and fault diagnosis of the gear box of wind turbine in wind farm, has important significance in improving the economic benefit of wind farm operation.

Gear box fault diagnosis, fault location is to determine the fault nature, and to determine the extent of failure, due to complex neural network with multi-mode and with the association, inference and memory function, which in recent years attracted a Diagnosis extensive research[5]. The only drawback is in the field of fault diagnosis, neural networks practical constraints the main factor is the lack of large representative sample of training. Because the number of equipment failure, after all, is limited by the rate of accumulation of such data, it is difficult to train "highly skilled" neural networks. The good news is the emergence of support vector-based theory of a small sample of the neural network learning is possible. In short, the gear fault diagnostic techniques and contemporary fusion of cutting-edge science is the gear fault diagnosis technology development, diagnostic technology becomes more

intelligent. However, the fault diagnosis of the measured sample data sets, and more balanced sample set is not the fault of the relatively small number of samples. How to use this data sample is not balanced better in the diagnosis, fault diagnosis is an urgent problem. With the extensive application, researchers have pointed out that the data does not bring balance to the classification of learning difficulties and challenges, foremost of which is significantly lower performance of classifiers.

The class imbalance problem is one of the relatively new problems that emerged when machine learning matured from an embryonic science to an applied technology, amply used in the worlds of business, industry and scientific research. Although practitioners might already have known about this problem early, it made its appearance in the machine learning data mining research circles about a decade ago[6]. Its importance grew as more and more researchers realized that their data sets were imbalanced and that this imbalance caused suboptimal classification performance. The class imbalance problem typically occurs when, in a classification problem, there are many more instances of some classes than others[7]. In such cases, standard classifiers tend to be overwhelmed by the large classes and ignore the small ones. In practical applications, the ratio of the small to the large classes can be drastic such as 1 to 100, 1 to 1,000, or 1 to 10,000 (and sometimes even more).

Nowadays, ensemble learning is becoming a hot topic in the machine learning and bioinformatics communities, which has been widely used to improve the generalization performance of single learning machines[8]. For ensemble learning, a good ensemble is one whose individuals are both accurate and make their errors on different parts of the input space [9]. The most popular methods for ensembles creation are Bagging and Boosting [10]. The effectiveness of such methods comes primarily from the diversity caused by re-sampling the training set. Random forest was also. Feature selection is used to produce different subsets for different learning machines.

Although the imbalanced data sets that classifier performance degradation, but feature selection can improve the performance of classifier. Feature selection refers to the concentration from the original feature selection makes some evaluation criteria of optimal feature subset. Its purpose is according to some criterion selected feature subset minimal; effect makes tasks such as classification, regression and feature selection before reaching approximate even better. By means of feature selection, feature and some irrelevant or redundant task is deleted, the simplified data sets will obtain more precise model, easier to understand.

Some research works in machine learning community has been made in classification problem on imbalanced data sets, which by Professor Zhou Zhihua, who proposed EasyEnsemble classifier, achieved better results than other methods [10]. In basis of EasyEnsemble classifier, feature selection used for classification problems on imbalanced data sets, we proposed based on PSO Particle Swarm Optimization (PSO) algorithm feature selection EasyEnsemble PSOEE (PSO based feature selection for EasyEnsemble).

The rest of this paper as follows: Part II briefly introduces the EasyEnsemble classifier, and then describes in detail the algorithm based on Particle Swarm Optimization based feature selection for EasyEnsemble PSOEE; the third section describes the algorithm used to test the UCI data sets and experimental settings; fourth part gear fault diagnosis carried out on experimental data sets; the fifth part of the text are summarized.

COMPUTATIONAL METHODS

The EasyEnsemble classifier is an under-sampling algorithm, which independently samples several subsets from negative examples and one classifier is built for each subset. All generated classifiers are then combined for the final decision by using Adaboost [10].
The pseudo-code of EasyEnsemble is rewritten as in Algorithm 1.

Algorithm 1 The EasyEnsemble algorithm

Input：Training data set, Number of individuals T
Output：Ensemble model N

1. Begin
2. for $k=1:T$
3. Generate a training subset S_{rk}^{-} from negative training set S_r^{-} by using the Bootstrap sampling technique, the size of S_{rk}^{-} is the same with that of S_r^{+}
4. Train the individual model N_k on the training subset $S_{rk}^{-} \cup S_r^{+}$ by using AdaBoost with weak classifiers $h_{k,j}$ and corresponding weights $\alpha_{k,j}$

$$N_k(x) = \mathrm{sgn}\left(\sum_{j=1}^{nk} \alpha_{k,j} h_{k,j}(x) - \theta_k\right)$$

5. End for
6. Ensemble the obtained models N like

$$N(x) = \mathrm{sgn}\left(\sum_{k=1}^{T} \sum_{j=1}^{nk} \alpha_{k,j} h_{k,j}(x) - \sum_{k=1}^{T} \theta_k\right)$$

7. End

Particle Swarm Optimization based feature selection for EasyEnsemble,Feature selection refers to pick out some of the most effective feature dimension in order to reduce the feature space from the original feature. High dimensional data contains many redundant features, even the noise characteristics; the existence of these features will not only greatly increase the training time and computational complexity of the algorithm, and may decrease the accuracy of classification. Therefore, feature selection in high dimensional data can be effective in removing irrelevant and redundant features so as to improve the efficiency of the learning algorithm to reduce the computational complexity [11].

Particle swarm optimization algorithm is originally attributed to Kennedy and Eberhart in 1995[12]. particle swarm optimization (PSO) is a computational method that optimizes a problem by iteratively trying to improve a candidate solution with regard to a given measure of quality. PSO optimizes a problem by having a population of candidate solutions, here dubbed particles, and moving these particles around in the search-space according to simple mathematical formulae over the particle's position and velocity. A basic variant of the PSO algorithm works by having a population (called a swarm) of candidate solutions (called particles). These particles are moved around in the search-space according to a few simple formulae. The movements of the particles are guided by their own best known position in the search-space as well as the entire swarm's best known position. When improved positions are being discovered these will then come to guide the movements of the swarm. The process is repeated and by doing so it is hoped, but not guaranteed, that a satisfactory solution will eventually be discovered[13].

The basic principle of particle swarm algorithm is based on the assumption that in a D target in the search space，There are m particles consist of a population, The i particles is expressed as a D dimensional vector $\vec{X}_i = (x_{i1}, x_{i2}, \cdots, x_{iD}), i = 1, 2, \cdots, m$, Position ofi particles in the D dimension of the search space is $\vec{X}_i$. $\vec{V}_i = (v_{i1}, v_{i2}, \cdots, v_{iD})$ Is the i particles flying speed， $\vec{P}_i = (p_{i1}, p_{i2}, \cdots, p_{iD})$ is the optimal position of the Iparticles so far to searchpbest， $\vec{P}_g = (p_{g1}, p_{g2}, \cdots, p_{gD})$ is the optimal location to the entire population gbest. The basic PSO optimization algorithm of particle updating equations is as follows：

$$v_{id}(k+1) = v_{id}(k) + c_1 r_1 (p_{id}(k) - x_{id}(k)) + c_2 r_2 (p_{gd}(k) - x_{id}(k)) \quad (1)$$

$$x_{id}(k+1) = x_{id}(k) + v_{id}(k+1) \ (i = 1, 2, \cdots, m; d = 1, 2, \cdots, D) \quad (2)$$

where，k is the iterative times；Learning factor c_1 and c_2 is a nonnegative constant，as 2 in the general； r_1 and r_2 are two random number that uniform distribution between [0,1]; $v_{id} \in [-v_{\max}, v_{\max}]$, $v_{\max}$ is a constant preset； $v_{id}(k)$ is the D dimensional of i velocity vector in the K iteration; is the D dimensional of i position vector in the K iteration； $p_{id}(k)$ is the D dimensional component of best position of i particle; $p_{gd}(k)$ is the D dimensional component of best position of population; Iterative conditional termination is set to the maximum number of iterations or (and) the best position of particle swarm to search to satisfy the minimum fitness value.

The performance evaluation on the basis of particle is the particle's fitness, so we must choose the appropriate fitness function. In this paper, the fitness of the population can understand the properties of the selected feature subset that represented as a particle. Performance evaluation criterion of feature subset are classic consistent measurement、accuracy measurement and classical measurement. In this paper the accuracy of measurement is used that which uses the correct classification rate to evaluate the performance of feature subset.

The i of fitness function is defined f (i): $f(i) = a(i) - p\frac{m_{ic}}{m_{all}}$, Where, a(i) is the correct rate estimation of classification in feature subset that selected by i; p is the classification accuracy and the number of selected features

of the equilibrium coefficient, p = 0. 2; mic is the number of sub feature selected for particle of i features; mall is the total number of features.

Based on particle swarm algorithm, we proposed PSO (Particle Swarm Optimization based feature selection for EasyEnsemble) for imbalanced data sets. At first, It use EasyEnsemble to get the integrated model, and then based on integrated model use particle swarm optimization algorithm get optimal feature subset by calculation of the training data set, and finally get the new integrated model using EasyEnsemble algorithm in the feature subset. The detailed algorithm is described as follows:

Algorithm 2 The PSOEE algorithm

Input：Training data set $S_r = \{(x, y)\}$, Balance coefficient P

Output：Ensemble model N

1. Begin
2. Train the ensemble model N_{temp} on the training set S_r by using EasyEnsemble.
3. Initializing the particle swarm in each particle parameters, using the formula (3) for fitness Calculate,to update the particle velocity and position, in the stop conditions to get the optimal feature subset
4. Generate the optimal training subset $S_{r-optimal}$ from S_r according to the above optimal features.
5. Re-train the model N on the optimal training subset $S_{r-optimal}$.
6. End

EXPERIMENTS ON UCI DATA SETS

To test the POSEE algorithm, five data sets selected from UCI machine learning repository[14]. 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. Information about these data sets are summarized in Table 1. where Size is the number of examples, Feature is the number of descriptors, #min/#max is the size of minor/major class, and Ratio is the size of major class divided by that of minor class.

In EasyEnsemble, 5 subsets are sampled, i.e. T is set to 5 during experiments, and on each an ensemble containing 15 weak learners are trained. Thus, the final ensemble generated by EasyEnsemble will contain 15*5=75 weak learners. In all experiments, we use the SVM with C = 1,σ = 10 as the weak learner.

To compare the results fairly, we use the 3-fold cross validation procedure. Using the top ranked features selected by a feature selection method, together with their expression values in the training dataset, one can build an EasyEnsemble that will decide for each testing example the class it belongs to. Only the expression values for those selected features in the testing example are used for such a decision making. This is a standard way to test the quality of those selected features, to examine how well the resulting classifier performs. Note that testing examples are not included in the training phrase.

TABLE 1The properties of the UCI data sets for comparison

Data set	Feature	Class	Size	Min/Max	Ratio
audio	70	24	226	48/178	3.71
voting records	16	2	435	168/267	1.59
proc_c	13	5	303	36/267	7.42
soy_a	34	19	307	40/267	6.68
backup	35	19	683	88/595	6.76

Since the class distribution of the used data set is skew, prediction accuracy (ACC) may be misleading. Therefore, AUC (Area Under the Curve of Receiver Operating Characteristic (ROC))[10] is used to measure the performance. To furthermore describe the different learning methods, we also define the various measures as below, where TP; TN;FP; FN, stand for the number of true positive, true negative, false positive, false negative samples at classification time, respectively. ACC, TPR(true positives ratio), TNR(true negatives ratio), and BAC (balanced accuracy) are defined as:

$TPR = TP / Pos$ ； $TNR = TN / Neg$ ； $ACC = (TP + TN) / N = (TP + TN) / (Pos + Neg)$ ； $BAC = (TPR + TFR) / 2$

It is compared between PSOEE algorithm and EasyEnsemble that does not use feature selection. The experimental results are shown in Table 2.

TABLE 2 Statistical results of AUC，BAC，TPR，TNR on the UCI data sets

Data set	All	PSOEE	All	PSOEE	All	PSOEE	All	PSOEE
	AUC		BAC		TPR		TNR	
audio	0.7973	0.8323	0.7913	0.8123	0.7757	0.8632	0.8328	0.8987
Voting_records	0.9451	0.9531	0.9521	0.9587	0.9523	0.9609	0.9586	0.9539
proc_c	0.6540	0.7418	0.6543	0.6989	0.6287	0.6418	0.6841	0.7746
soy_a	0.9113	0.9309	0.9139	0.9377	0.9315	0.9573	1.000	1.000
backup	0.9549	0.9801	0.9746	0.9768	0.9569	0.9629	1.000	1.000
Average	0.8525	0.8876	0.8572	0.8769	0.8490	0.8772	0.8951	0.9254

From Table 2, the results can be seen that AUC value that the feature is not selected is 0.8525, AUC values of the PSOEE algorithm is 0.8876, which is 4.11% higher than the former; not to feature selection TPR value is 0.8572, TPR of PSOEE algorithm value is 0.8769, which is 3.32% higher than the former; the TNR values not to feature selection is 0.8490, TNR value of PSOEE algorithm is 0.8772, which is 3.38% higher than the former. Experimental results on UCI datasets show that, on the imbalanced data set, PSOEE algorithm does improve generalization ability and improve the value of AUC, TPR, TNR.

Wind turbines are running in the tens of meters altitude, wind disturbances affected by mechanical transmission, the load change is more complex, especially in our part of the wind farms in mountainous or hilly areas, affected by topography airflow distortion, so wind turbines are complex long-term cross-working under variable loads due to wind uncertainties rotational speed wind turbine gearbox changing the internal structure of complex vibration signal usually in the form of AM and FM, superimposed on each other coupling between the components, resulting in signal space distribution characteristics disorganized, signals have non-stationary, uncertainty and complexity, etc. [15], which are making wind turbine vibration signal analysis more complicated.

In the study of wind turbine gearbox failure, due to objective conditions, it is difficult to collect large amounts of failure modes in short-term. Fault simulation is a good experimental research methods, it is artificial manufacture of certain failures under certain conditions in the gear box to simulate reality of some failure modes, and then through the analysis, thus to determine and validate fault diagnosis.

Gear fault data sets used in this paper is from a gearbox fault simulation system dragging along with a motor. This system includes hub, drive shaft, gear boxes, bearings. The gearbox fault simulation system can simulate several type of faults without damaging the physical structure. Gearbox fault simulation system simulate the four gearbox fault (tooth crack failure, shaft unbalance, shaft misalignment, axial movement), then the signal data collected use the methods of the time-domain , amplitude domain analysis, parameter extraction amplitude domain and frequency domain parameters as training samples. 6 characteristic parameters are selected (the peak factor, kurtosis, pulse index, margin index, power spectrum entropy and correlation dimension) as a wind turbine gearbox fault diagnosis eigenvalues. The sampling frequency is 5120Hz, sampling points for each sample point data is 8192. 10 group's data in each failure are selected as training data samples. 40 fault samples are collected to merge the related fault data, see table 3:

TABLE3 Description of the data set of gearbox

Data set	Feature	Class	Size	Min/Max	Ratio
Gearbox	6	4	40	10/30	3.00

From TABLE 4, the results can be seen that AUC value that the feature is not selected is 0.9402, AUC values of the PSOEE algorithm is 0.9817, which is 4.41% higher than the former; not to feature selection TPR value is 0.9532, TPR of PSOEE algorithm value is 1.0000, which is 4.91% higher than the former; the TNR values not to feature selection is 0.8910, TNR value of PSOEE algorithm is 0.9573, which is 7.44% higher than the former. BAC is not the value of feature selection is 0.9445; BAC value of PSOEE algorithm is 0.9789, which is 3.64% higher than the former. Experimental results on UCI datasets show that, on the gear box data set, PSOEE algorithm does improve generalization ability and improve the value of AUC, TPR, and TNR. From comparison of the results of TNR and TPR it can be seen, TPR improved more values in PSOEE algorithms. It means for improving the prediction accuracy of a small class of the sample is the main reason to improve the AUC in the imbalanced data set.

TABLE 4 Statistical results of AUC，BAC，TPR，TNR on gearbox data set

Data set	All	PSOEE	All	PSOEE	All	PSOEE	All	PSOEE
	AUC		BAC		TPR		TNR	
Gearbox	0.9402	0.9817	0.9445	0.9789	0.9532	1.0000	0.8910	0.9573

Granular computing is a simulation of human global analysis capability. The computer's processing speed is far greater than the speed of the human brain, but the computer is not as intelligent as human beings. This is mainly because humans have a very strong global analysis capability to turn complex problems into a relatively simple model from a variety of different size or level. Granular computing is a computational paradigm of information processing, covering all the granularity related theory, methods, techniques and tools. Most of the researches on granular computing are theoretical research, and less in the applications. The analysis, discussion and application of granular computing model to specific problems are an urgent need for the study.

CONCLUSION

This paper presents a novel algorithm PSOEE (PSO based feature selection for EasyEnsemble)to solve the gear box fault diagnosis data that is not balance, while using the AUC data classification indicators as uneven performance evaluation criteria, the final gear in the UCI data sets and data sets on a fault experiment. Experimental results show that PSOEE algorithm improves the classification prediction accuracy of the data set. Imbalanced data set classification problem is one of the problems of data classification, and its main difficulty is the uneven data set of their own characteristics and limitations of the traditional classification algorithms result. PSOEE algorithm is through remove redundant features to improve the classification performance that is a solution to uneven data sets an effective classification method. PSOEE characteristics of the algorithm the original feature set a subset of the proportion of total direct impact on the results of the algorithm, which will work in the future to do further research.

Acknowledgments
This work was supported by Shanghai Municipal Natural Science Foundation (No. 14ZR1417200), Innovation Program of Shanghai Municipal Education Commission (No. 14YZ157), National Natural Science Foundation of China(No. 61374136).

REFERENCES

[1] He Dexin, *Engineering Science***2011**,6,95-100.
[2] The European Wind Energy Association,Wind energy—the facts http://www.ewea. Org/index.php?id=91.**2014**
[3] Yang, WX; Tavner, PJ; Crabtree, CJ; Feng, Y; Qiu, Y,*Wind Energy*, **2014**,17(5),673-693.
[4] Gray, C.; Watson, S,*Wind Energy*, **2010**, 13(5), 395–405.
[5] Tavner PJ; Xiang JP; Spinato F,*Wind Energy*,**2007**,10, 1–18.
[6] Wasikowski, M; Xue-wen Chen,,*IEEE Transactions on Knowledge and Data Engineering*,**2010**,22(10), 1388-1400.
[7] Xiaowan Zhang; Bao-Gang Hu, I*EEE Transactions on Knowledge and Data Engineering*, **2014**,26(12) ,2872-2885.
[8] Galar, M; Fernández, A; Barrenechea, E,*IEEE Transactions on Systems, Man, and Cybernetics*, **2012**,42(4), 463-484.
[9] M. You; J.-M. Liu; Y. Chen; G.-Z. Li,*International Journal of Computational Intelligence Systems*,**2012**,5(4),668-678.
[10] X.-Y. Liu; J. Wu; Z.-H. Zhou,*Proceedings of International Conference on Data Mining***2006**, 965-969
[11] Bing Xue; Mengjie Zhang; Browne, W.N,*IEEE Transactions on Cybernetics*, **2013**,43(6) 1656-1671.
[12] Kennedy J; Eberhart R,*Proceeding of IEEE International Conference on Neutral Networks Australia*,**1995**,1942-1948.
[13] Goh C K;Tan K C;Liu D S,*Journal of Operational Research*,**2010**,202,42-45.
[14] Blake C; Keogh E; Merz C J,1998 UCI repository of machine learning databases. http://www.ics.uci.edu/mlearn/ MLRepository.html. Department of Information and Computer Science, University of California, Irvine, California,. 12 Jan **2014**
[15] Feng, YH; Qiu, YN;Crabtree, CJ,*Wind Energy*, **2013**, 16(5), 728-740.