2014

A GA-SVM Hybrid Classifier for Multiclass Fault Identification of Drivetrain Gearboxes

Dingguo Lu
University of Nebraska-Lincoln, stan1860@huskers.unl.edu

Wei Qiao
University of Nebraska-Lincoln, wqiao@engr.unl.edu

Energy Conversion Congress and Exposition (ECCE), 2014 IEEE
Year: 2014
Pages: 3894 - 3900, DOI: 10.1109/ECCE.2014.6953930

A GA-SVM Hybrid Classifier for Multiclass Fault Identification of Drivetrain Gearboxes

Dingguo Lu and Wei Qiao
Power and Energy Systems Laboratory
Department of Electrical Engineering
University of Nebraska-Lincoln
Lincoln, NE 68588-0511 USA
stan1860@huskers.unl.edu, wqiao@unl.edu

***Abstract*—This paper presents a genetic algorithm (GA)-support vector machine (SVM) hybrid classifier for multiclass fault identification of drivetrain gearboxes in variable-speed operational conditions. An adaptive feature extraction algorithm is employed to effectively extract the features of gearbox faults from the stator current signal of an AC machine connected to the gearbox. The multiclass GA-SVM classifier is used to identify the faults in the gearbox according to the fault features extracted. A GA is designed to find the optimal parameters of the SVM to obtain the best classification accuracy. The proposed hybrid classifier is validated on a gearbox connected with a permanent-magnet synchronous machine with three different faults. Experimental results show that the multiple types of gearbox faults can be effectively identified and classified by the proposed hybrid classifier with better accuracy than the traditional SVM classifier.**

***Keywords*—Adaptive resampling, classification, condition monitoring, fault diagnosis, drivetrain gearbox, genetic algorithm (GA), support vector machine (SVM)**

I. INTRODUCTION

Identifying gearbox faults in a timely manner is of great importance to ensure safe operation of the drivetrains of many rotating machinery systems, such as wind energy systems, electric vehicles, aircrafts, electric trains, etc. [1]-[4]. Current-based methods have been demonstrated effective in identification of drivetrain gearbox faults, and their advantages over the conventional vibration-based methods have been discovered in terms of cost, implementation, reliability, accessibility, and robustness.

In [5] a current-based method was proposed to identify the fault types for drivetrain gearboxes running at variable speeds. That method used an adaptive feature extraction algorithm to effectively extract the features of gearbox faults from the stator current signal of the permanent-magnet synchronous machine (PMSM) connected to the gearbox. A multiclass radial basis function (RBF)-kernel support vector machine (SVM) classifier was designed to identify different faults in the gearbox according to the fault features extracted. That study examined the possibility of diagnosing different gearbox faults by adopting the SVM model.

The accuracy and stability of the SVM classifier for fault identification rely on its parameter setting. To design an effective SVM classifier, the values of the parameters of the SVM have to be chosen properly. In the previous work [5], the best parameters of the SVM classifier were found by a trial-and-error method, which cannot guarantee the parameters obtained to be optimal. An automatic parameter optimization method is therefore required to obtain the optimal design of the SVM. Conventionally, the problem of optimal parameter setting for a SVM classifier was solved by using the grid search method [6]. However, the grid-searching method is expensive in terms of computational cost and data requirement. Consequently, the purpose of this study is to propose an algorithm that is capable of effectively finding the optimal parameters of the SVM classifier to yield the highest classification accuracy and generalization ability for identifying the gearbox faults, while requiring lower computational cost and data requirement.

GA is an evolutionary computation technique commonly used to solve complex linear and nonlinear optimization problems. It simulates the natural evolution process of a population to explore the problem space through mutation, crossover, and selection operations applied to the individuals of the population. GA is an effective, parallel and global optimization approach and, therefore, is a good candidate to find the optimal parameters of the SVM classifier in this paper.

This paper proposes a GA-SVM hybrid classifier for multiclass fault identification of a drivetrain gearbox operating in variable-speed conditions. A three-step feature extraction algorithm is employed to extract the frequency-domain features of gearbox faults from the generator stator current measurements. A multiclass RBF-SVM classifier is designed to identify the faults of the gearbox according to the

This work was supported in part by the U.S. National Science Foundation under grant ECCS-1308045.

fault features extracted. The parameters of the SVM classifier are optimized by a GA method. Experimental studies are carried out to validate the proposed method for identification of multiple types of gear faults in a drivetrain gearbox.

II. Multiclass Fault Identification of Drivetrain Gearboxes

A. Characteristic Frequencies of Drivetrain Gearboxes

In a drivetrain gearbox, due to the torsional vibrations induced by transmission errors in the input, pinion, and output wheels and the stiffness variation of gear tooth contact, rotational and meshing frequencies appear in the torque signature of the output shaft [7]. If the drivetrain gearbox is connected to an AC electric machine, the characteristic frequencies of the gearbox vibration will modulate the stator currents of the AC machine and generate sidebands across the dominant components in the frequency spectra of the currents [8]. As demonstrated in [5], [8], [9], the faults in a gearbox that change the characteristic frequencies of the gearbox vibration will subsequently change the sideband distribution in the frequency spectra of the current signals of the AC machine. For a drivetrain consisting of a multistage gearbox connected with a PMSM, the characteristic frequencies of the faults in a gearbox vibration signal include the shaft rotation-related frequencies f_i ($i = 1, 2, \ldots, I$, I is the number of gears in the gearbox), and gear meshing-related frequencies $f_{mesh,j}$ ($j = 1, 2, \ldots, J$, J is the number of gear pairs in the gearbox). As a result, the frequencies $f_{sideband}$ of the sideband components in the stator currents can be expressed in the following form:

$$f_{sideband} = kf_s \pm \sum_{i=1}^{I} l_i f_i \pm \sum_{j=1}^{J} m_j f_{mesh,j} \qquad (1)$$

where f_s is the fundamental frequency of the stator current, and $k, l_i, m_j = 0, 1, 2, \cdots$.

B. Adaptive Feature Extraction

The characteristic frequencies of gearbox faults in the PMSM stator currents are related to gearbox shaft speeds, and become nonstationary when the shaft speeds vary with time. In order to identify the faults in the gearbox operating in nonstationary conditions, an adaptive feature extraction algorithm was developed in [5] to extract the fault signatures from the nonstationary PMSM stator current signals for pattern recognition in following steps. The proposed feature extraction algorithm consists of three main steps, as shown in Fig. 1, where f_1, f_2 and f_3 are the rotating frequencies of gearbox shafts, and the details can be found in [5], where PSD stands for power spectrum density.

First, an adaptive signal resampling algorithm illustrated in Fig. 2 is developed to resample the original current samples with a fixed sampling rate using an adaptive sampling rate such that the nonstationary characteristic frequencies of faults, become constant.

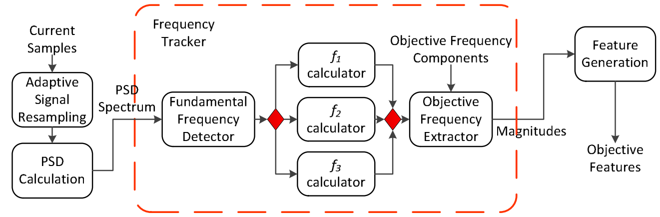

Fig. 1. Schematic diagram of the feature extraction scheme.

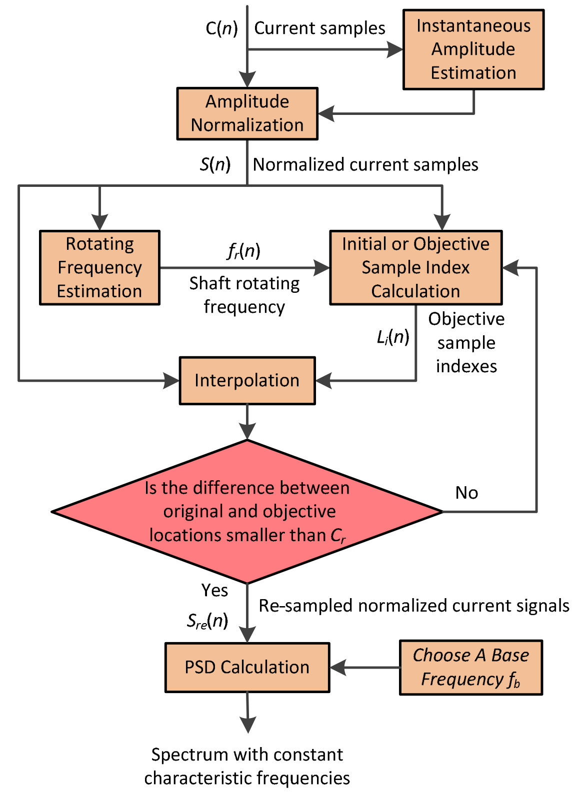

Fig. 2. Schematic of the adaptive signal resampling algorithm.

Second, a frequency tracker is designed to extract the objective frequency components with their magnitudes from the resampled signals according to the mathematical relationships among the rotating speeds of the gearbox gears and shafts. In this work, the objective frequency components are the sidebands around the fundamental frequency of the resampled stator current.

Third, a feature generator is designed to generate the features related to faults. Two groups of sidebands around the fundamental frequency are calculated using the frequency tracker. The first group is the sidebands caused by the first to fourth orders of the gearbox input shaft rotating frequency, which are expressed as $f_s \pm pf_1$, where $p = 1, 2, 3$ and 4. The second group is the sidebands caused by the three gearbox shaft rotating frequencies, which are expressed as $f_s \pm f_1$, $f_s \pm f_2$ and $f_s \pm f_3$. The magnitudes of these two groups of sidebands are first normalized with respect to the magnitude of the fundament frequency component. Two features are then generated for each group from the normalized sideband magnitudes: the standard deviation and summation as in (2) and (3), respectively.

$$SD = \left[\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2\right]^{1/2} \quad (2)$$

$$M = \sum_{i=1}^{N} x_i \quad (3)$$

where x_i $(i = 1, 2, \dots, N)$ is the normalized average magnitude of the ith sideband pair, $\bar{x}$ is the mean value of the normalized magnitudes of all sideband pairs, and N is the number of sideband pairs being processed.

C. Multiclass SVM Classifier

SVM is an inherent candidate for binary classification. The basic idea of a binary SVM classifier is to construct a hyperplane as the decision plane, which separates the positive (+1) and negative (-1) classes with the largest margin. As shown in Fig. 3, $\boldsymbol{\omega} \cdot x + \mathbf{b} = 0$ represents the hyperplane. The margin is the sum of the distances from the hyperplane to the closest data points of each of the two classes. These closest data points along the boundaries expressed by $\boldsymbol{\omega} \cdot x + \mathbf{b} = \pm 1$ are defined as Support Vectors (SVs). In the SVM, a nonlinear kernel function is used to map the input vectors into a higher-dimensional feature space through some nonlinear separating hyperplane and, thus, makes the data linearly separable in the feature space although the original input vectors are nonlinearly separable in the input space, as illustrated in Fig. 4. This paper uses a RBF kernel to construct the SVM.

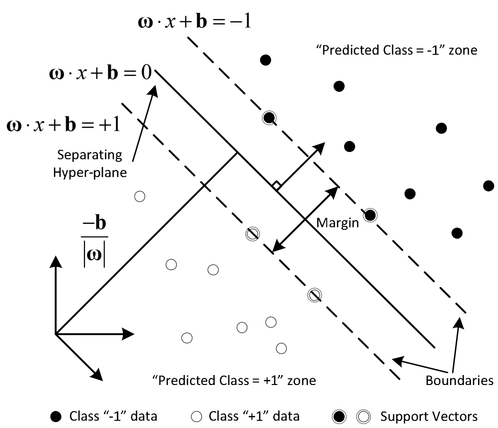

Fig. 3. SVM classification.

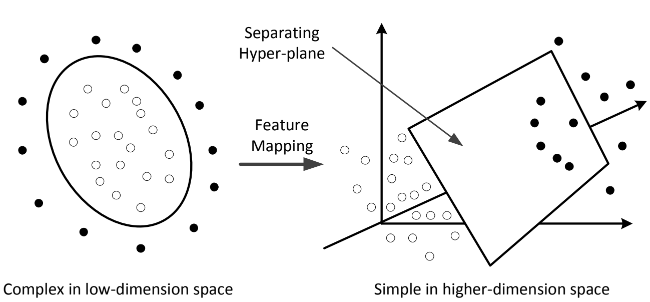

Fig. 4. Feature mapping by using the kernel function.

Additionally, since the fault identification of drivetrain gearboxes is a multiclass problem, effort has been made to extend the SVM for multiclass classification. The concept of multiclass classifier is to construct multiple binary SVM classifiers under certain strategies [10]. The critical issue is to choose an appropriate classification strategy. This work utilizes a One-Against-One (OAO) strategy [11], which has advantages of using a small number of training samples for each classifier, symmetric data structure, and low computational loads.

In summary, a multiclass RBF-SVM-based classifier with an OAO multiclass classification strategy is designed to evaluate the condition of the gearbox. The SVM uses the features generated by the adaptive feature extraction scheme described in Section II-B and outputs coded fault types.

III. OPTIMIZATION OF THE SVM CLASSIFIER VIA GA

The selection of the SVM parameters is critical to the performance of the SVM. The RBF-SVM has the following parameters to choose [10]:

1) Regularization parameter C, which determines the tradeoff between the fitting error of the SVM model and the model complexity. It is also called the cost parameter of the SVM.

2) Kernel free parameter σ, which determines the bandwidth of the RBF kernel and, subsequently, defines the nonlinear mapping from the input space to the high-dimensional feature space.

The GA is used to determine the optimal values of C and γ that assure the optimal predication accuracy and generalization ability simultaneously.

B. Principles of GA

As one of the evolutionary computation techniques, GA is an adaptive optimization methodology modeling the process of natural selection. It has been successfully applied to optimization and machine learning problems [12].

A GA evolves a population of chromosomes as potential solutions to an optimization problem, and the optimal solution is obtained after a series of iterative operations. In general, the initial population is generated randomly. In this paper, the GA is used to find the optimal values of the two parameters, C and σ, of the SVM. Therefore, each chromosome comprises two parts $X = [b_C, b_\sigma]$ and is constructed by using a binary coding system, as shown in Fig. 5. Once initialized, the GA implements a generational evolution strategy and employs the following genetic operations to search for the optimal solution for the problem: fitness evaluation, parent selection, crossover, and mutation.

(1) Fitness Evaluation: A fitness function is designed to assess the performance of each chromosome in each iteration. In this research, the goal of the search strategy is to maximize the accuracy of the SVM classifier for identification of the gearbox faults. Therefore, the GA is designed to find the values of C and σ that maximize the

fitness function value, i.e., the classification accuracy of the SVM.

(2) Parent Selection: This step is to choose appropriate parents for reproduction. The proposed GA-SVM classifier adopts the heuristic method with a generation gap [12] to add experience to the population. The generation gap determines the rate of individuals to be selected. In this study, a generation gap of 0.9 is used. It means that 90% of the population with the best fitness values is kept for the following crossover and mutation operation.

(3) Crossover: The crossover operation allows new solution regions in the problem space to be explored. During the crossover operation, a selected chromosome is crossed over at one or more positions randomly assigned. The resulting chromosomes are then combined with the rest chromosomes to generate the offspring. The generated offspring replaces the old population to form a new population in the next generation. The process of crossover operation is shown in Fig. 6. In this study, the probability of crossover occurring between pairs of individuals is set as 0.7.

(4) Mutation: In a binary coding system, during mutation, "1" and "0" are mutated to each other. Each element of a chromosome is mutated with a given probability, which is 0.7 in this study. The mutation process is shown in Fig. 6 as well.

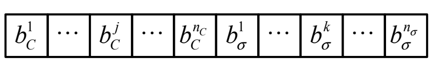

Fig. 5. Encoding in a chromosome, where n_C and n_σ are the numbers of binaries for C and σ, respectively.

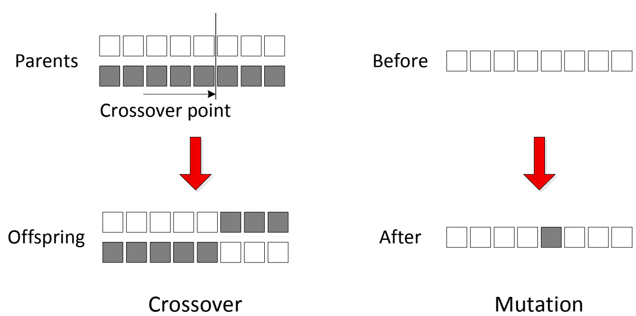

Fig. 6. The crossover and mutation operations.

C. Proposed GA-SVM Hybrid Classifier

A GA-SVM hybrid classifier is then proposed in this study for multiclass fault identification of drivetrain gearboxes. The GA is used to find the optimal values of C and σ that assure the optimal predication accuracy and generalization ability of the SVM simultaneously. The flowchart of the proposed hybrid classifier is shown in Fig. 7. The GA tries to search for the global optimal solution to enable the SVM to fit selected datasets. The SVM classifier with the optimal parameters is then used for gearbox fault identification.

It should be pointed out that the values of C and σ are limited into certain ranges, to assure the generalization capability of the SVM. In this work, the ranges are set to be $C \in [0.1, 100]$ and $\sigma \in [0.1, 1000]$.

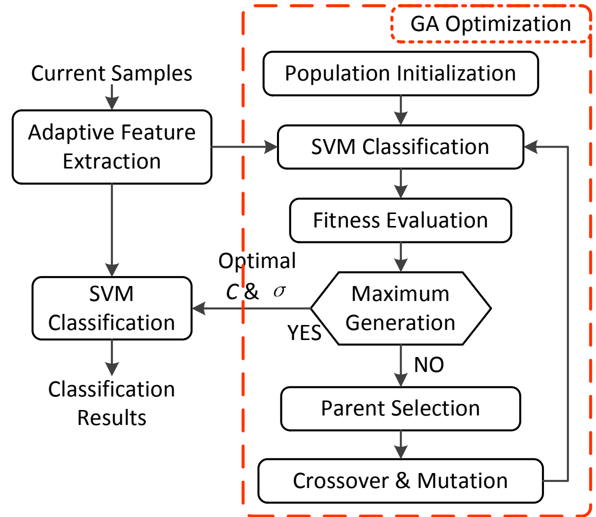

Fig. 7. The proposed GA-SVM hybrid classifier.

IV. EXPERIMENTAL VALIDATION

A. Experiemetnal System

The proposed hybrid classifier is validated for diagnosis of multiple types of faults in a test drivetrain gearbox. Fig. 8 shows the experimental system setup, which consists of a 300-W PMSM driven by a variable-speed induction motor (IM) through two back-to-back connected gearboxes. These two gearboxes are two-stage gearboxes with 3 shafts and 4 gears. Their internal structure is shown in Fig. 9. The test gears are mounted at the input shaft of the test gearbox and pretreated by artificially generating various faults which were commonly observed in industrial systems, including one-tooth missing, two-tooth missing, and a gear crack (Fig. 10). One phase stator current of the PMSM is recorded with a sampling rate of 10 kHz for 210 seconds, during which the rotating speed of the PMSM is varied randomly in the range of 297 to 891 rpm. A total number of 1326 data records are collected, including 340 healthy cases, 318 one-tooth-missing cases, 293 two-tooth-missing cases, and 375 gear crack cases.

The two-stage drivetrain gearbox considered in this work includes 3 frequency components that are related to the shaft rotation and 2 frequency components that are related to gear meshing. The latter is difficult to observe in the signals, since the magnitudes of these frequency components are largely attenuated due to the damping effect of the mechanical system. Hence, for the drivetrain studied in this work, (1) becomes

$$f_{sideband} = kf_s \pm l_1 f_1 \pm l_2 f_2 \pm l_3 f_3 \qquad (4)$$

Fig. 8. The experimental system.

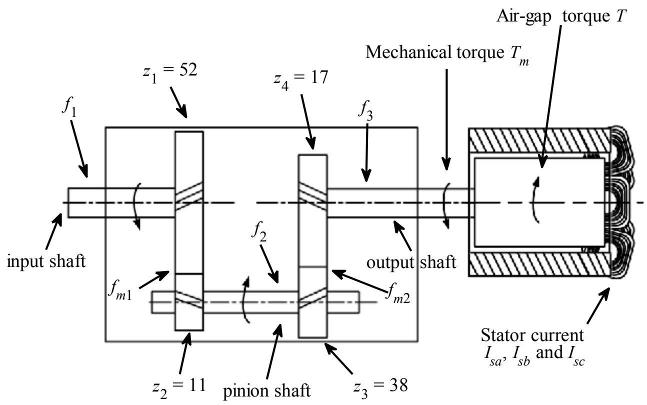

Fig. 9. Structure of the test drivetrain consisting of a two-stage gearbox connected to a PMSG with characteristic vibration frequencies.

B. Feature Set Preprocessing for SVM Classifier

The proposed adaptive feature extraction algorithm is first applied to the raw PMSM stator current samples to generate the fault features, as discussed in Section II. A total number of 1326 sets of fault features, which are the statistical standard deviation and summation of the magnitudes of the sidebands in the current spectra described in Section II, are thus obtained from the complete datasets. The extracted fault features are then used as the inputs of the proposed GA-SVM hybrid classifier, where the GA is executed to optimize the parameters of the multiclass SVM classifier.

As described in Section III, the implementation of the proposed GA-SVM hybrid classifier requires two sets of samples. One set is used to find the optimal values of the SVM parameters using the GA, and the other set is used for to validate the designed GA-SVM classifier for gearbox fault identification. Therefore, the original dataset is divided into two groups prior to applying the proposed method. In this work, the feature sets obtained is divided randomly and evenly into two groups. As a result, the first group of the dataset for parameter optimization has 664 feature sets, including 159 for the baseline case, 159 for the one-tooth-missing case, 147 for the two-tooth-missing case and 187 for the gear crack case; and the second group of dataset for GA-SVM classifier validation has 662 feature sets, including 170 for the baseline case, 159 for the one-tooth-missing case, 146 for the two-tooth-missing case, and 187 for the gear crack case. Once the optimal values of the parameters are obtained, SVM classification can be carried out to verify the proposed method following the similar training and testing procedure as in [5].

C. Mulitclass Fault Identification

The proposed GA-SVM classifier is then validated using the obtained feature sets. The parameters of the GA should also be carefully selected prior to executing the proposed method, as discussed in Section III. These parameters determine the GA's generalization capability, convergence speed, and computational complexity. The parameters of the GA used in this work are listed in Table I.

In this paper, the multiclass GA-SVM classifier is designed to identify four different conditions of the test gearbox, including one healthy case and three faulty cases. Therefore, the output of the classifier should be coded to represent these four cases. Table II illustrates the output codes of the proposed classifier.

TABLE I
GA PARAMETER SETTING

Population Size	20	C_{max}	100	C_{min}	0.1
Maximum Generation	100	σ_{max}	1000	σ_{min}	0.1
Generation Gap	0.9	Crossover Probability	0.7	Mutation Rate	0.7

TABLE II
OUTPUT CODING IN PROPOSED MULTICLASS SVM CLASSIFIER

Condition	Healthy	One-tooth-missing	Two-tooth-missing	Gear crack
Code	0	1	2	3

In the proposed method, the GA is first used to search the optimal values of the SVM parameters, starting with a randomly generated population. The classification results over 100 generations of GA execution are shown in Fig. 11. The evaluation result of the classification accuracy in each generation is shown in Fig. 11. The classification accuracy starts with a relatively low value of approximately 96% and then increases as more generations evolve over time. The classification accuracy converges to a steady-state value after the 10th generation. After that the classification accuracy is not a constant value, but fluctuates with a period of about 20 generations as the populations evolve. The maximum accuracy is 99.1% while the minimum accuracy is 97.4% with a mean value of 98.4%. That means the GA is searching around the best set of $[C, \sigma]$ over the generations. The best

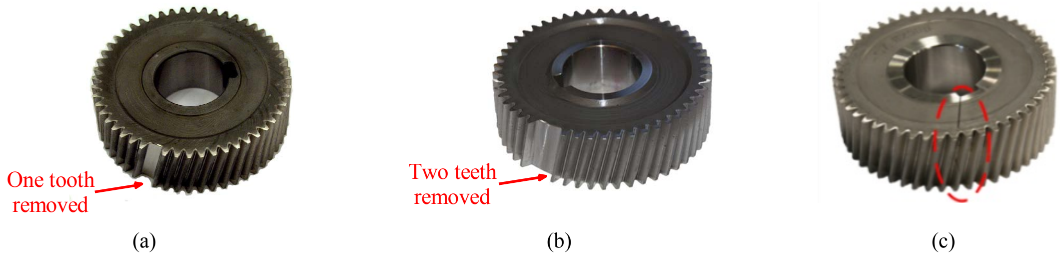

Fig. 10. The test gears with (a) one-tooth-missing, (b) two-tooth-missing, and (c) gear crack.

accuracy achieved is 99.1% with the optimal SVM parameters $C_{opt} = 0.89493$ and $\sigma_{opt} = 3.0079$, which are then used to design the SVM classifier to identify the faults using the second group of feature sets for validation of the proposed method.

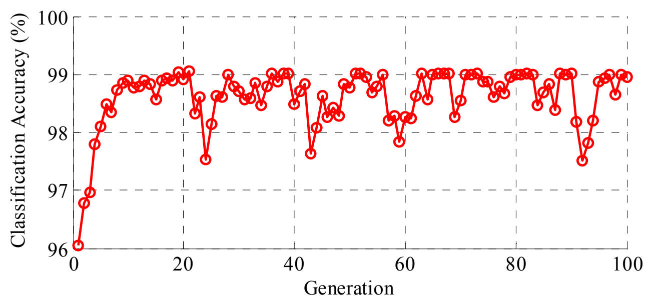

Fig. 11. Classification accuracy of the GA-SVM hybrid classifier over generations.

Monte Carlo studies are carried out to testify the generalization capability of the proposed method. The Monte Carlo method is a computational algorithm that uses a repeated random sampling process to obtain numerical results for the distribution of an unknown probabilistic entity. It is designed for the systematic investigations of the performance of statistical tools under various conditions and, therefore, is useful for evaluating the performance of the proposed method. The division of the original dataset introduces the difference of a classification problem. The similar process of data division and GA-SVM classifier design and validation is carried out in the Monte Carlo studies for 100 times to obtain the probabilistic distribution of the classification accuracy of the proposed classifier.

The statistics of the classification results in Monte Carlo studies are summarized in Table III. The results clearly indicate that the multiple types of faults in the drivetrain gearbox are properly classified by the proposed method consistently. Meanwhile, Compared to the SVM classifier proposed in [5], the proposed GA-SVM hybrid classifier has higher classification accuracy and is more stable (i.e., a smaller standard deviation), as indicated by Table III.

V. Conclusions

Gearbox is one of the most important components in the drivetrains of rotating machinery systems widely used in the industry. The faults in a drivetrain gearbox can make the system offline for a substantially long time, leading to increased operation and maintenance cost.

A novel GA-SVM hybrid classifier has been designed for multiple fault identification of drivetrain gearboxes operating in variable-speed conditions. The proposed hybrid classifier is a multiclass RBF kernel-SVM classifier with an OAO classification strategy, where the parameters of the SVM were optimized by a GA to maximize the fault classification accuracy. The resulting SVM classifier identifies the faults in the gearbox according to the fault features extracted by an adaptive feature extraction algorithm. Experimental studies have been performed on a drivetrain gearbox test bed with multiple types of faults. The results have proved that the proposed GA-SVM hybrid classifier could effectively identify the faults in the gearbox, with higher classification accuracy and stability than the traditional SVM classifier.

TABLE III
CLASSIFIER CLASSIFICATION ACCURACIES

	Max	Min	Average	Standard deviation
SVM	96.25%	87.52%	91.5%	0.0242
GA-SVM	99.77%	96.1%	98.7%	0.0083

References

[1] K. Smolders, H. Long, Y. Fend, and P. Tavner, "Reliability analysis and prediction of wind turbine gearboxes," in *Proc. European Wind Energy Conf.*, Apr. 2010, pp. 162-165.

[2] G. Maggetto and J. Van Mierlo, "Electric and electric hybrid vehicle technology: a survey," in *Proc. IEE Seminar Electric, Hybrid and Fuel Cell Vehicles*, Apr. 2000, pp. 1/1-111.

[3] A. Boglietti, A. Cavagnino, A. Tenconi, and S. Vaschetto, "The safety critical electric machines and drives in the more electric aircraft: A survey," in *Proc. 35th Annual Conf. of the IEEE Industrial Electronics Society*, Nov. 2009, pp. 2587-2594.

[4] H. W. Lee, K. Kim, and J. Lee, "Review of maglev train technologies," *IEEE Trans. Magnetics*, vol. 42, no. 7, pp. 1917-1925, Jun. 2006.

[5] D. Lu and W. Qiao, "Adaptive feature extraction and SVM classification for real-time fault diagnosis of drivetrain gearboxes," in *Proc. IEEE Energy Conversion Cong. and Expo.*, Sept. 2013, pp. 3934-3940.

[6] C. Hsu, C. Chang, and C. Lin, "A practical guide to support vector classification," National Taiwan University, Taipei, Taiwan, Tech. Rep., Apr. 2010.

[7] S. H. Kia, H. Henao, and G. A. Capolino, "Gearbox monitoring using induction machine stator current analysis," in *Proc. IEEE Int. Symp. on Diagnostics for Electric Machines, Power Electronics and Drives*, Sept. 2007, pp. 149-154.

[8] D. Lu, X. Gong, and W. Qiao, "Current-based diagnosis for gear tooth breaks in wind turbine gearboxes," in *Proc. IEEE Energy Conversion Cong. and Expo.*, Sept. 2012, pp. 3780-3786.

[9] D. Lu and W. Qiao, "Frequency demodulation-aided condition monitoring for drivetrain gearboxes," in *Proc. IEEE Transportation Electrification Conference & Exposition*, Jun. 2013, pp. 1-6.

[10] S. R. Gunn, "Support vector machines for classification and regression," School of Electronics and Computer Science, University of Southampton, Tech. Rep., May 1998.

[11] C. Hsu and C. Lin, "A comparison of methods for multiclass support vector machines," *IEEE Trans. Neural Networks*, vol. 13, no. 2, pp. 415-425, Mar. 2002.

[12] K. S. Tang, K. F. Man, S. Kwong, and Q. He, "Genetic algorithms and their applications," *IEEE Signal Processing Magazine*, vol. 13, no. 6, pp. 22-37, Nov. 1996.