Conditional entropy-based feature selection for fault detection in analog circuits

Selección de características basadas en entropía condicional para la detección de averías en circuitos analógicos

Ting Long[1], Shiqi Jiang[1], Hang Luo[2,3] and Changjian Deng[1]

[1] School of Control Engineering, Chengdu University of Information Technology, Chengdu, 610225, Sichuan, China,leamonlong@hotmail.com
[2] School of Manufacturing Science and Engineering, University of Sichuan, Chengdu, 610065, Sichuan, China, luohang2002@163.com
[3] School of Computer Science, McGill University, Montreal, H3A0E9, Quebec, Canada

DOI: http://dx.doi.org/10.6036/7920 | Recibido: 31/12/2015 • Aceptado: 30/03/2016

RESUMEN

- Para detectar defectos paramétricos en circuitos analógicos, se propone en el presente estudio un nuevo algoritmo de selección de características basado en la entropía condicional, que se integra en un proceso de detección de defectos basado en una máquina de vector soporte (SVM). Para prevenir la pérdida significativa de actuaciones eficaces se ejecutó un proceso de muestra con una frecuencia significativamente alta. Para reducir el sobrecalentamiento de la computación, el algoritmo de selección de características basado en la entropía condicional se introdujo posteriormente para comprimir los vectores brutos de observación en vectores nuevos de observación. La entropía condicional se usó para actualizar la probabilidad condicional de un defecto basado en una nueva información de defecto que eventualmente hace más clara la probabilidad de defecto. Aplicando el algoritmo propuesto de selección de características, se pueden comprimir los datos brutos más certeramente y maximizar la información, escogiendo las dimensiones que deben incluirse en los nuevos vectores de observación. Los resultados de la simulación mostraron que el proceso de detección de defectos presentado en el estudio podría clasificar el espacio vector de proceso no lineal de circuitos analógicos y conseguía una proporción menor de malas clasificaciones que otros métodos actuales (p. ej. El método equidistante y el método basado en la probabilidad convencional).
- **Palabras clave:** Detección de defectos, Entropía condicional, Máquina de vector soporte (SVM), Clasificación.

ABSTRACT

To detect parametric faults in analog circuits, a novel support vector machine (SVM)-based fault detection approach, which was integrated with conditional entropy-based feature selection algorithm, was proposed in this study. In preventing the significant loss of the effective features, a sampling process was executed with a significantly higher frequency. The side effect of this process showed that raw observation vectors were of extremely high dimensions. To reduce computation overhead, the feature selection algorithm based on conditional entropy was put forward to compress raw observation vectors into new observation vectors. The conditional entropy was used to update the conditional probability of a fault based on new fault information, which eventually made the fault probability more clear. By applying the proposed feature selection algorithm, it can compress raw data more wisely, and maximize the information in choosing the dimensions to be included in the new observation vectors. Simulation results showed that the fault detection approach presented in the study could classify non-linear feature vector space of the analog circuits. and achieved a lower misclassification rate than other current methods (i.e., equidistant method and conditional probability-based method).

Keywords: Fault detection, Conditional entropy, Support vector machine (SVM), Classification.

1. INTRODUCTION

In mixed integrated circuits (ICs) and systems on a chip (SOCs), the scales of digital circuits are usually larger than those of analog circuits. However, fault detection of analog circuits is more difficult because of the circuits' continuous signals in input ports and output ports, in fault model, in tolerances of analog components, and in other parts.

In analog circuits, a fault means any change in the value of a component. It can cause deviation from the normal behavior of the circuit under test (CUT). A parametric fault refers to a tolerable deviation of an analog system from the normal specifications [1]; it negatively affects the system's performance but does not result in system termination [2]. Fault detection is performed to discover deviations in a circuit. Fault detection for analog circuits is difficult, especially for parametric faults. Parametric faults deviate very subtle. Such subtlety of parametric faults results in highly mixed raw feature space, which necessitates non-linear classification and feature selection.

Parametric faults are detected by observing system outputs. Methods proposed in [3-5] are used to detect parametric faults in analog electronic circuits. They selected optimal test frequencies for fault detection. However, no all-purpose test frequency selection method can be applied to all kinds of parametric faults.

The raw system output observations of parametric faults may involve excessive irrelevant information and noise, incurring largely unnecessary computation and making fault detection infeasible. Therefore, raw system output observations or raw observation vectors are usually abstracted into features, that is, feature vectors, for efficient fault detection. This process is called feature selection and reduces computation overhead but may cause information loss. After feature selection, the remaining information must be sufficient for high-precision fault detection. Most studies on feature selection focus on maximizing the information in

choosing dimensions to be included in the feature vectors. Some feature selection methods were proposed in [6-8] for statistical analysis and fault detection. However, whether these methods are applicable in analog circuit fault detection is uncertain.

A fault model of analog circuits based on circle equation was proposed in [9]. Tian et al. [9] used complex fault model to build features. However, this model can be only used in linear analog circuits. Starzyk et al. [10] used entropy to select test points, but their method was not used for fault detection. A fault detection method [11] using entropies as inputs was only tested in switched current circuits. Two fault detection methods based on support vector machine (SVM) was proposed in [12] and [13]. They built feature vectors with feature extraction based on principal component analysis (PCA). Feature extraction is used to build new features from original features, for example, wavelet transformation transforms time domain feature vectors to frequency domain feature vectors [14, 15]. However, new features in feature extraction may lose the visual meaning of the original features. In analog circuit fault detection, original features have better intuitive, physical, or visual meanings. Therefore, feature selection algorithms are used to keep the visual meaning of the original features in fault detection. A conditional entropy-based feature selection algorithm proposed in this paper can keep the visual meaning of the original features with minimizing information loss.

Pan et al. [16] proposed an equidistant feature selection method in the context of a linear classification fault detection methodology. For example, a raw observation vector would become a feature vector with an equidistance of 5. This method is easy to implement and runs fast. However, it does not consider using information theory to evaluate the information content of each raw observation vector dimension. It may cause information loss. After feature selection, the remaining information must be sufficient for high precision fault detection. The problem is maximizing the information in choosing the dimensions to be included in the feature space. To address this problem, we propose a conditional entropy-based feature selection algorithm and evaluate this algorithm in SVM-based fault detection, which is a mainstream fault detection method for parametric faults. For the rest of the paper, the term "fault" shall implicitly refer to "parametric fault" unless otherwise denoted.

The rest of the paper is organized as follows: Section 2 describes the SVM-based fault detection and proposes our conditional entropy-based feature selection algorithm. Section 3 evaluates the performance of our proposal and contrasts our algorithm with other feature selection algorithms. Section 4 concludes the paper.

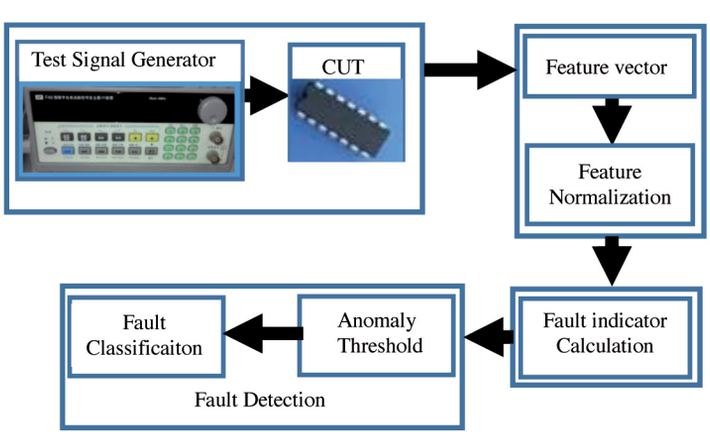


Fig. 1: Overview of the proposed framework for fault detection

2. METHODOLOGY

2.1. FAULT DETECTION FOR PARAMETRIC FAULTS

An overview of the proposed framework for fault detection is shown in Fig. 1. In our work, a sinusoidal signal is used as the test signal. Fault detection based on SVM for parametric faults involves two phases: the training phase and the diagnosis phase. The training phase consists of three steps (see Fig. 2).

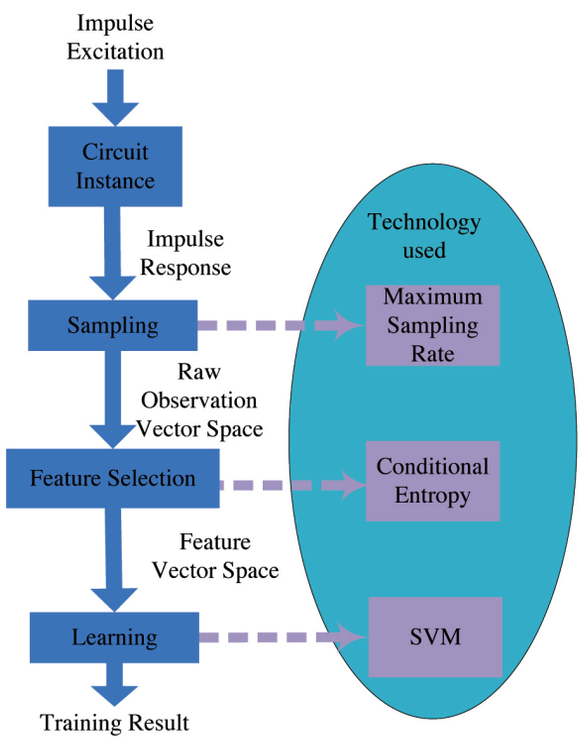


Fig. 2: Steps of training fault detection SVM

As mentioned in Section 1, Step 1 (Sampling): The raw system outputs are observed by sampling the impulse responses from the training circuit instances (a benchmark set of known normal and faulty circuit instances). Each raw observation is a vector of n samples, that is, these raw observation vectors belong to an n-dimensional vector space called the raw observation vector space.

Step 2 (Feature Selection): Feature selection is conducted, mapping the raw observation vectors of n-dimension into feature vectors of m-dimension, where $m < n$. The m-dimensional vector space is called the feature vector space.

Step 3 (Learning): The fault detection SVM is trained with the feature vectors obtained in Step 2.

Fig. 2 shows the technology used in each of the three steps. In particular, Step 1 samples the impulse responses of the training circuit instances with the maximum sampling rate possible to retrieve raw observation vectors. Step 2 conducts conditional entropy-based feature selection to compress the n-dimensional raw observation vectors into m-dimensional feature vectors. Step 3 builds the fault detection SVM by learning the feature vectors derived in Step 2. These three steps are further explained in the rest of this section.

The diagnosis phase involves similar steps:

Step 1 (Sampling): This step is the same as Step 1 of the training phase. However, instead of sampling with a maximal rate on a training circuit instance (which we know a priori whether it is normal or faulty), a test circuit instance whose state (normal or faulty) is unknown is sampled. The result is also an n-dimensional raw observation vector.

Step 2 (Feature Selection): This step is also the same as Step 2 of the training phase. The same feature selection algorithm is used to map the n-dimensional raw observation vector into the m-dimensional feature vector.

Step 3 (Detection): This step is straightforward. The feature vector derived in Step 2 is plugged into the SVM learned in Step 3 of the training phase. If the SVM indicates that the feature vector lies in the "normal" area of the feature vector space, then the circuit instance under test is normal; otherwise, the circuit instance under test is faulty.

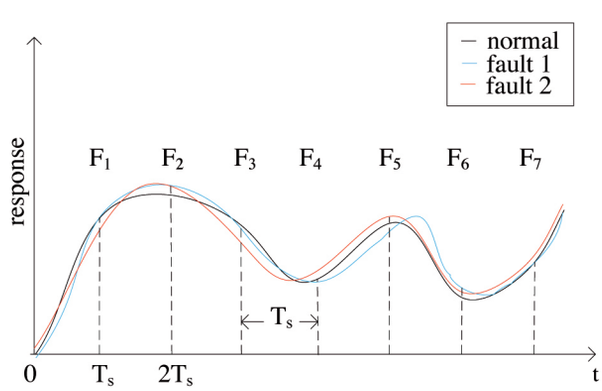
To be concise, in the following sections, we shall focus on describing the training phase of our SVM-based fault detection methodology because of the similarity of the training phase and the diagnosis phase. The three steps of the training phase are elaborated in Sections 2.1, 2.2, and 2.3.

2.2. SAMPLING FOR BUILDING RAW OBSERVATION SPACE

A parametric fault is detected through the observation of system outputs. However, parametric faults are usually difficult to detect. If circuit responses are under-sampled, then useful features may be missed.

Although the renowned Nyquist sampling theorem states that sampling at more than twice the bandwidth of a signal guarantees no loss of information, knowing the bandwidth of its response signals a priori is difficult if a circuit has potential parametric faults. Therefore, all that can be done is to always use the maximum sampling rate allowed by our observation device.

Fig. 3 illustrates a sampling process with a loss of the effective features. It illustrates three response examples, which include a normal response, a response of fault 1, and a response of fault 2. Each sampling point with the sampling frequency Fs is a feature of a response. The feature is labeled Fi. For example, the first feature of the normal response is , and the second feature of the normal response is . Therefore, the feature vector of the normal response is (, ,...,). In a similar manner, the feature vectors of the responses of fault 1 and fault 2 in Fig. 3 are represented by (, ,...,) and (, ,...,) separately. We use 'l_fea' to represent the length of the feature vectors, and l_fea= Fs/BW, where BW is the bandwidth of the CUT. The features may be similar in the different responses of the circuit instances, such as and , and . If the features of the fault response are similar with those of the normal response, then the features are useless for the classification in fault diagnosis. This similarity means that the effective features are lost. If the feature vectors include many useless features, then achieving effective fault diagnosis using these feature vectors would be difficult.



$T_s=1/F_s$ where F_s means sampling frequency.

Fig. 3: Sampling process with loss of effective features

When oversampling technology is used, the sampling frequency is significantly higher than twice the bandwidth of the response. This relationship means that the sampling process is executed with a significantly higher frequency, and more features can be included in a feature vector. It prevents the loss of as many effective features as possible.

The response signal is impulse response of each circuit instance in this study. The impulse response is used because it is the simplest response that represents the fault.

2.3. FEATURE SELECTION ALGORITHMS

The side effect of the maximum sampling rate strategy in Step 1 (see Section 2.1) shows that raw observation vectors are of extremely high dimensions. For example, for a sampling device capable of a 100 MHz sampling rate (common to today's digital analyzers), observing a response signal for only 1.2 microseconds will result in a 120-dimensional raw observation vector. Moreover, when observations are made at the granularity of seconds, the raw observation vector dimensions will result in the magnitude of 10^8.

Such high dimensional raw observation vectors will incur unaffordable computation in SVM learning (training phase Step 3, see Section 2.3). In addition, for most raw observation vectors, most of their dimensions do not carry meaningful information. Therefore, we have to reduce the dimensions of the raw observation vector space. For each raw observation vector, only the elements for the chosen dimensions are retained; all other elements are discarded. This step maps each raw observation vector into a *feature vector* of a lower dimension. This process is called *feature selection*, and the reduced dimension vector space is called the *feature vector space*.

Although feature selection reduces computation and storage overhead, it may cause information loss [17]. After feature selection, the remaining information must be sufficient for high precision (i.e., *efficient*) fault detection. The problem lies on maximizing the information in choosing the dimensions to be included in the feature vector space [18]. In the following sections, two feature selection algorithms are proposed to address this problem: a conditional probability- and a conditional entropy-based algorithm.

Given a CUT, we generate training circuit instances by varying component values within the allowed ranges or by injecting faults into the CUT. Hence, some generated circuit instances are normal, whereas others are faulty. The raw observation vector of a normal circuit instance is labeled *"normal,"* whereas that of a faulty circuit instance is labeled *"faulty."* We let X represent whether an element of one raw observation vector is close to normal: $X = 1$ means that the element is close to normal, and $X = -1$ means that the element is close to faulty. We let Y represent whether the raw observation vector where X belongs to is "normal" or "faulty." $Y = 1$ means that the raw observation vector is "normal," and $Y = -1$ means that the vector is "faulty."

With the abovementioned context, feature selection means that the information gain of each dimension in the raw observation vector space needs to be evaluated in classifying "normal" and "faulty" training vectors, and the most information-rich dimensions to be retained in the feature vector space need to be selected.

2.3.1. Conditional Probability Algorithm

In probability theory, conditional probability is one of the most fundamental and important concepts. When an event results from (by assumption, presumption, assertion, or evidence) another

event, a conditional probability can be used to measure the probability. We take the conditional probability P(*A*|*B*) as an example. P(*A*|*B*) denotes the probability of event *A* given event *B*. In information theory, conditional probability is used to denote an update of the probability of an event based on new information. In test engineering, conditional probability is commonly used to update the probability of a fault based on new fault information. It makes the fault probability clearer.

Conditional probability can be calculated with joint probability. Joint probability is easily confusable with conditional probability. Joint probability denotes the probability of two events in conjunction. We take the joint probability P(*A*, *B*) as an example. P(*A*, *B*) denotes the probability of both events *A* and *B* together. It can also be written as P(*AB*).

P(*A*|*B*) is used to examine the probability of event *A* as it is restricted to event *B*. It means that the probability of event *A* is measured when event *B* has or will have occurred. The conditional probability P(*A*|*B*) can be measured by joint probability with P(*A*, *B*) multiplied by P(*B*), where P(*B*) is the probability of event *A*, and P(*A*, *B*) is the joint probability of events *A* and *B*. P(*A*|*B*) is the updated P(*A*) after evidence *B* is accounted for.

In the preliminary algorithm of the last section, event *X* represents that an element of one raw observation vector is close to normal or faulty. The possibility of *X* is measured by observation vectors under a normal or faulty circuit. Event *Y* represents that the CUT is normal or faulty. The possibility of *Y* is measured by learning.

With the abovementioned context, feature selection means the information gain of each dimension in the raw observation vector space needs to be evaluated in classifying "normal" (normal) and "faulty" (faulty) training vectors, and the most information-rich dimensions to be retained in the feature vector space must be selected.

In choosing the most information-rich raw observation vector space dimensions, a straightforward heuristic method is used to find those dimensions with high probabilities of $(Y=1\,|\,X=1)$ and $(Y=-1\,|\,X=-1)$ and low probabilities of $(Y=1\,|\,X=-1)$ and $(Y=-1\,|\,X=1)$. Therefore, a simple formulation would be (in case of selecting only one dimension in the feature vector space) finding the dimension in the raw observation vector space whose *X* optimizes the following objectives:

$$\max\{p(Y=1\,|\,X=1)+p(Y=-1\,|\,X=-1)\} \tag{1}$$

and

$$\min\{p(Y=1\,|\,X=-1)+p(Y=-1\,|\,X=1)\} \tag{2}$$

where $p(\cdot)$ is the probability function.

We note that Objective Function (2) can be transformed into

$$\begin{aligned}&\min\{p(Y=1\,|\,X=-1)+p(Y=-1\,|\,X=1)\}\\&=\min\{1-p(Y=-1\,|\,X=-1)+1-p(Y=1\,|\,X=1)\}\\&=\min\{2-p(Y=-1\,|\,X=-1)-p(Y=1\,|\,X=1)\}\\&=\min\{2-(p(Y=-1\,|\,X=-1)+p(Y=1\,|\,X=1))\}\\&=\max\{p(Y=-1\,|\,X=-1)+p(Y=1\,|\,X=1)\}.\end{aligned} \tag{3}$$

Therefore, Objective Function (2) equals Objective Function (1). Hence, we only need to focus on Objective Function (1).

We note that to select *m* dimensions in the feature vector space, we only need to apply the optimization algorithm *m* times, and each time removes the selected optimal dimension from the candidate set.

2.3.2. Conditional Entropy Algorithm

However, we notice that a better formulation of our feature selection problem is a multiple objective optimization problem:

$$\begin{aligned}&\max\{p(Y=1\,|\,X=1)\}\ \text{and}\ \max\{p(Y=-1\,|\,X=-1)\}\ \text{and}\\&\min\{p(Y=-1\,|\,X=1)\}\ \text{and}\ \min\{p(Y=1\,|\,X=-1)\}\end{aligned} \tag{4}$$

among all dimensions in the raw observation vector space. To make this multiple objective optimization problem solvable, we approximate the problem by merging the multiple objectives into one. Objective Function (1), the preliminary algorithm's formulation, is only one straightforward way of merging. Whether a better way of merging the objectives exists, which retains more information contained in individual conditional events $(Y=1\,|\,X=1)$, $(Y=-1\,|\,X=-1)$, $(Y=-1\,|\,X=1)$, and $(Y=1\,|\,X=-1)$, is unclear.

We notice that conditional entropy [19] defined in information theory can accurately measure the information contained in a parameter for decision making. This note inspires us to use conditional entropy as the metric for feature selection. Specifically, according to information theory, we define the following conditional entropy function:

$$H(Y\,|\,X)=H(Y=1\,|\,X=1)+H(Y=-1\,|\,X=1)+H(Y=-1\,|\,X=-1)+H(Y=1\,|\,X=-1) \tag{5}$$

where

$$\begin{aligned}&H(Y=1\,|\,X=1)=-p(X=1,Y=1)\log_2 p(Y=1\,|\,X=1),\\&H(Y=-1\,|\,X=1)=-p(X=1,Y=-1)\log_2 p(Y=-1\,|\,X=1),\\&H(Y=-1\,|\,X=-1)=-p(X=-1,Y=-1)\log_2 p(Y=-1\,|\,X=-1),\\&H(Y=1\,|\,X=-1)=-p(X=-1,Y=1)\log_2 p(Y=1\,|\,X=-1).\end{aligned}$$

Our conditional entropy-based feature selection algorithm shall choose those dimensions with the least conditional entropy $H(Y\,|\,X)$ values. The intuition is stated as follows.

We see that conditional entropy is a non-negative real number. Intuitively, it considers two aspects of using $X = x$ to determine $Y = y$. First, $-\log_2 p(Y=y\,|\,X=x)$ measures the uncertainty of claiming $Y = y$ based on the observation of $X = x$. As $p(Y=y\,|\,X=x)$ increases, the uncertainty decreases. When $p(Y=y\,|\,X=x)=1$, the uncertainty reaches a minimum of 0. Second, the multiplier $p(X=x,Y=y)$ further weighs the uncertainty metric of $-\log_2 p(Y=y\,|\,X=x)$. If $p(X=x,Y=y)$ is larger, then the event $(X=x,Y=y)$ happens more often; hence, the uncertainty metric of $-\log_2 p(Y=y\,|\,X=x)$ is practically more useful (hence more weight), and vice versa.

In fact, information theory has proven that conditional entropy is the best metric to measure the uncertainty [10, 20] of using parameter *X* to determine *Y*. With this concept, we expect the conditional entropy-based feature selection to be a better approximation of the multiple objective optimization problem of Objective Function (4).

Thereafter, the feature selection based on the conditional entropy works in two steps:

1) Equation (5) is used to calculate the conditional entropy for each dimension of the raw observation vector space.

2) The *m* dimensions with the smallest conditional entropy value are chosen to be the dimensions selected from the feature vector space.

A remaining technical detail shows that a specific dimension of a given raw observation vector is deciding whether it maps to *X* = 1 (normal) or *X* = −1 (faulty).

Specifically, we let $C_{training}$, C_{normal}, and C_{faulty} be the whole training set of circuit instances, the subset of all "normal" training

circuit instances, and the subset of all "faulty" training circuit instances, respectively. We let $f^c = (F_1^c, F_2^c, ..., F_i^c, ..., F_n^c)$ be the raw observation vector for a specific circuit instance $c \in C_{training}$, where *n* is the dimension of raw observation vector space.

We now suppose that we are interested in evaluating the *i*th dimension of the raw observation vector space. Therefore, we need to map each F_i^c $(c \in C_{training})$ into either $X_i^c = 1$ or $X_i^c = -1$. This requirement is accomplished in four steps:

Step 1: The average value of $F_i^{\forall}$ for all $\forall \in C_{normal}$ is calculated with

$$\overline{F}_i^{normal} = \frac{1}{|C_{normal}|} \sum_{\varsigma \in C_{normal}} F_i^{\varsigma} \tag{6}$$

This value is regarded as the expected value of the *i*th dimension of a raw observation vector when the given circuit instance is normal.

Step 2: The deviation of F_i^c from $\overline{F}_i^{normal}$ is calculated by

$$\sigma_i^c = \frac{\left| F_i^c - \overline{F}_i^{normal} \right|}{\overline{F}_i^{normal}} \tag{7}$$

Step 3: The average deviation of each *normal* and *faulty* class is calculated by Equations (8) and (9), respectively:

$$\sigma_i^{normal} = \frac{1}{|C_{normal}|} \sum_{\varsigma \in C_{normal}} \frac{\left| F_i^{\varsigma} - \overline{F}_i^{normal} \right|}{\overline{F}_i^{normal}} \tag{8}$$

$$\sigma_i^{faulty} = \frac{1}{|C_{faulty}|} \sum_{\varsigma \in C_{faulty}} \frac{\left| F_i^{\varsigma} - \overline{F}_i^{normal} \right|}{\overline{F}_i^{normal}} \tag{9}$$

Step 4: The X^c value for F_i^c is decided by the threshold θ_i:

$$\theta_i = \frac{1}{2}(\sigma_i^{normal} + \sigma_i^{faulty}) \tag{10}$$

If $\sigma_i^c < \theta_i$, then $X_i^c = 1$; otherwise, $X_i^c = -1$.

2.4. LEARNING WITH SVM

After performing feature selection, we map the training raw observation vectors to a lower dimension feature vector space. Such a reduction of raw data allows us to carry out the last step of the training phase: learning with SVM.

The main idea of SVM is finding a separating hyperplane for the "normal" class and "faulty" class in the feature vector space (see Fig. 4). A separating hyperplane *S* is defined by a vector *w* (in feature vector space) and an offset *b* stated as follows:

$$S = \{f \mid \langle w, f \rangle + b = 0\} \tag{11}$$

where *f* is a vector in the feature vector space, and $\langle \cdot, \cdot \rangle$ means inner product.

The goal is finding the optimal *w* and *b*. SVM aims to find the separating hyperplane that maximizes the distance from the nearest training feature vectors [21, 22]. These nearest training feature vectors are called *support vectors*. SVM theory proves that for optimal *w* and *b*, hyperplane $\{f \mid \langle w, f \rangle + b = 1\}$ and $\{f \mid \langle w, f \rangle + b = -1\}$, which are parallel to the separating hyperplane $S = \{f \mid \langle w, f \rangle + b = 0\}$, shall intersect with all class "normal" and class "faulty" support vectors, respectively (see Fig. 3). Furthermore, the distance from any support vector to the separating hyperplane *S* is $1/|w|$.

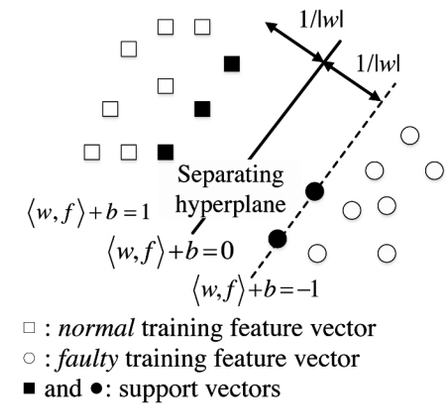


Fig. 4: SVM separating hyperplane

We suppose that *b* is known; thus, *w* must satisfy the following optimization problem

$$\begin{aligned} \min \quad & \frac{1}{2}\|w\|^2 = \frac{1}{2}\langle w \cdot w \rangle, \\ \text{s.t.} \quad & y_i(\langle w \cdot f_i \rangle + b) - 1 \geq 0, \quad i \in N_f \end{aligned} \tag{12}$$

where y_i denotes whether the training feature vector f_i belongs to the class of "normal" ($y_i = 1$) or "faulty" ($y_i = -1$). N_f is the index set of all training feature vectors. Optimization problem (12) can be solved by the following optimization problem on variable set $\{a_i\}$:

$$\begin{aligned} \min \quad & \frac{1}{2} \sum_{i,j \in N_f} a_i a_j y_i y_j \langle f_i, f_j \rangle - \sum_{i \in N_f} a_i \\ \text{s.t.} \quad & a_i \geq 0 \qquad i \in N_f \\ & \sum_{i \in N_f} y_i a_i = 0 \end{aligned} \tag{13}$$

where $\{a_i\}$ are also called Lagrangian multipliers. After solving Optimization Problem (13), we can derive *w* with $w = \sum_{i \in N_f} a_i y_i f_i$. Furthermore, if $a_i > 0$, then the corresponding f_i is a support vector. We can use this (theoretically, any of such) support vector to find *b*, as $\langle w, f_i \rangle + b = 1$ when f_i is a class "normal" support vector, and $\langle w, f_i \rangle + b = -1$ when f_i is a class "faulty" support vector.

The above SVM hypeplane algorithm is a linear classifier. SVM can also create nonlinear classifiers by applying kernel functions. Therefore, Optimization Problem (13) becomes

$$\begin{aligned} \min \quad & \frac{1}{2} \sum_{i,j \in N_f} a_i a_j y_i y_j K(f_i, f_j) - \sum_{i \in N_f} a_i \\ \text{s.t.} \quad & a_i \geq 0 \qquad i \in N_f \\ & \sum_{i \in N_f} y_i a_i = 0 \end{aligned} \tag{14}$$

where *K* is the kernel function. We can use similar approaches to derive the non-linear optimal separating hyperplane $S_{nonlinear}$:

hyperplane $S_{nonlinear}$:

$$S_{nonlinear} = \{f \mid \sum_{i \in N_f} a_i y_i K(f_i, f) + b = 0\} \tag{15}$$

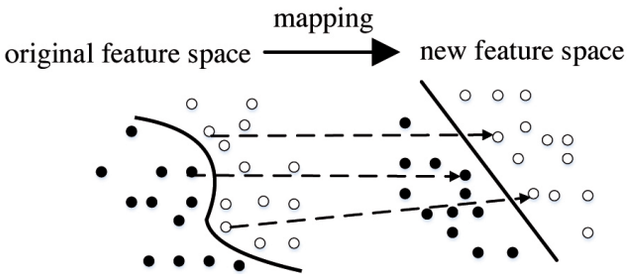


Fig. 5: Feature space of SVM

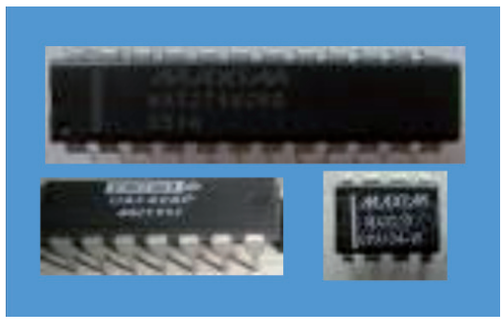


Fig. 6: Single integrated circuit comprising active filter elements

Intuitively, a kernel function non-linearly maps the original feature vector space into a new feature vector space (see Fig. 5). This method is often useful when training feature vectors (in the original feature vector space) are too intermingled to allow a linear separation hyperplane. In such cases, kernel function may map these training feature vectors into a new feature vector space where a linear separation hyperplane exists (see Fig. 5). This step allows linear classification in the new feature vector space, and the classification result corresponds (through the kernel function) to a non-linear classification in the old feature vector space.

Various kinds of kernel functions exist for SVM learning, for example, linear, polynomial, *radial basis function* (RBF), and sigmoid, and we can design our own kernel functions. Different kernel functions lead to different shapes of separating hyperplanes and hence different classification boundaries. Given a training set, a particular kernel function can possibly work well, whereas another kernel function may not. Until today, no good ways can be used to choose or design an optimal kernel function. In most cases, the choice or design depends on experience.

3. EXPERIMENTAL TEST AND RESULT ANANLYSIS

In analog integrated circuits, active filters are common. Single integrated circuits comprising active filter elements are shown in Fig. 6. Without the loss of generality, the active filter circuits in Fig. 7 are used to evaluate the performance of our fault detection methodology. For each CUT, we assign normal and faulty parameters to its components, to generate 100 normal and 100 faulty circuit instances as training set and another 100 normal and 100 faulty circuit instances as testing set. Without loss of generality, non-linear SVM learning described by (14) and RBF kernel function, which is a widely used non-linear kernel function for non-linear SVM learning, are used. In classification algorithm based on non-linear SVM misclassification rate can indicate fault detection accuracy.

Fig. 8 to Fig. 10 show the performance of fault detection with conditional entropy-based feature selection for CUT in Fig. 5. We use five different samples for training or testing. Fig. 8 shows the results for the two-pole active filter CUT (see Fig. 7(a)), Fig. 9 for the three-pole active filter CUT (see Fig. 7(b)), and Fig. 10 for the five-pole active filter CUT (see Fig. 7(c)). Each solid point of triangle or square in Fig. 8 to Fig. 10 represents a sample. The raw observation vectors are all of 120-dimension. We carry out three different trials of feature selection, reducing the dimension from 120 to 60, 40, 30 and 20.

Fig. 8 to Fig. 10 compare the accuracy of conditional probability-based feature selection and conditional entropy-based feature selection. The performances of all CUTs show that conditional entropy-based feature selection achieves better accuracy than conditional probability-based feature selection. Fig. 8 to Fig.

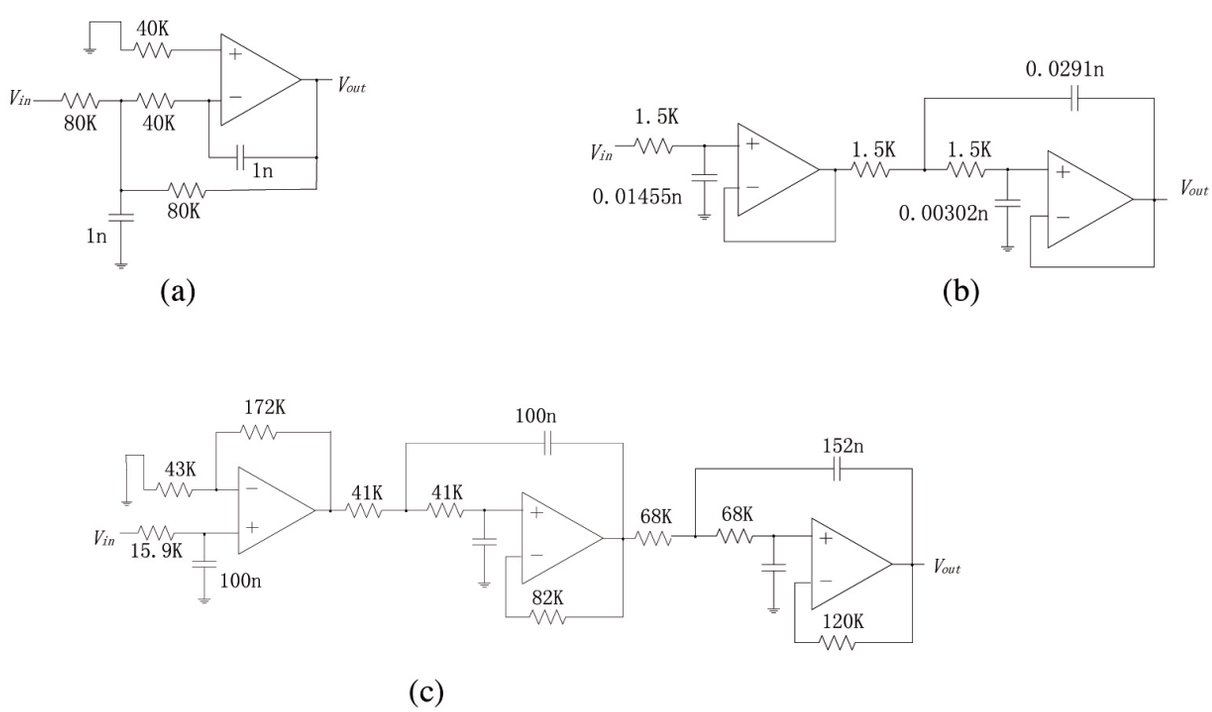


Fig. 7: Example CUT .(a) two-pole active filter. (b) three-pole active filter. (c) five-pole active filter

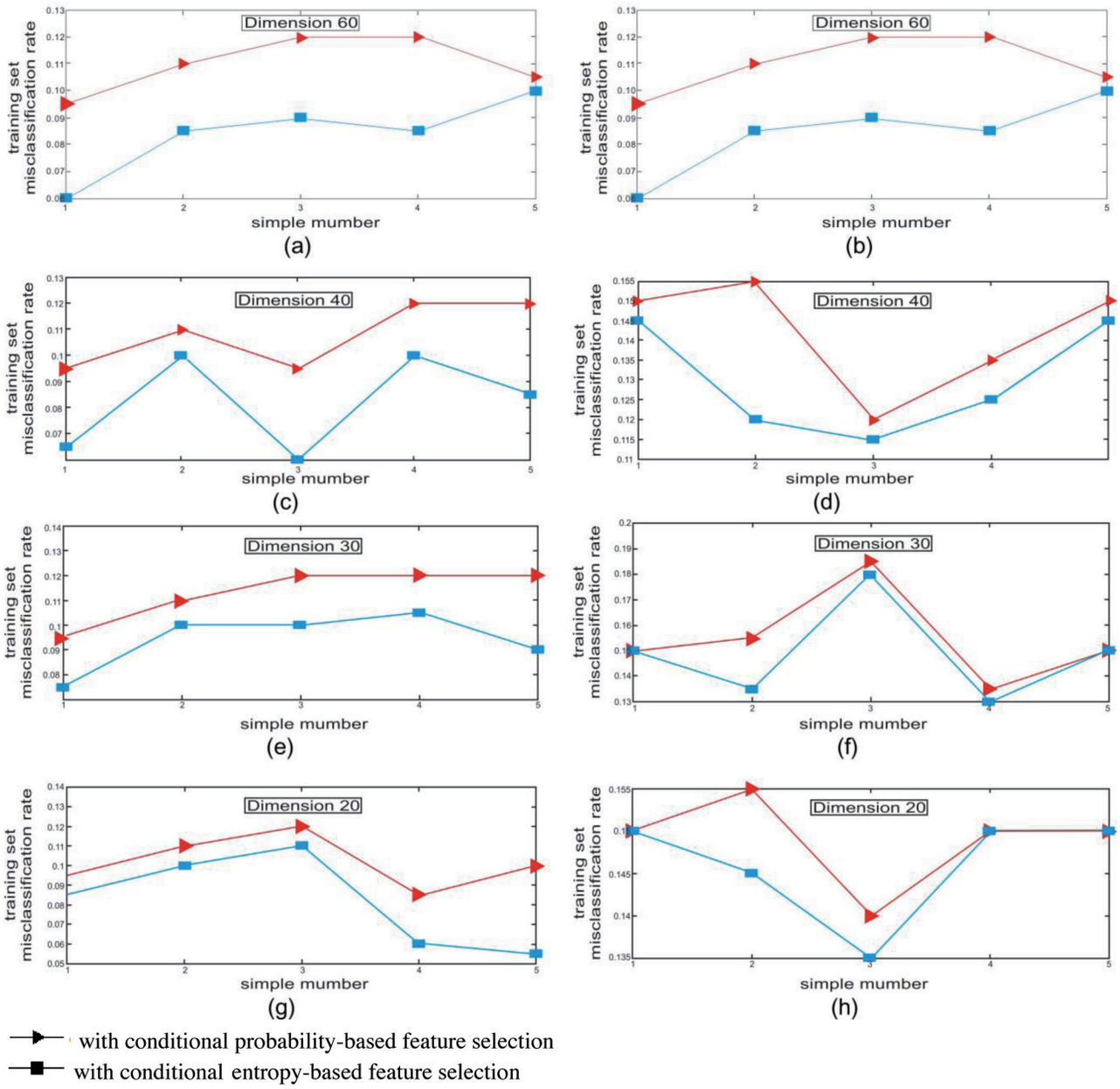


Fig. 8: Misclassification rate of fault detection with conditional probability-based and conditional entropy-based for two-pole active filter CUT. (a) training set of feature dimension 60. (b) testing set of feature dimension 60. (c) training set of feature dimension 40. (d) testing set of feature dimension 40. (e) training set of feature dimension 30. (f) testing set of feature dimension 30. (g) training set of feature dimension 20. (h) testing set of feature dimension 20

10 show that conditional entropy-based feature selection better evaluates the information content of each raw observation vector dimension when selecting features. The figures also illustrate that the conditional entropy-based feature selection solves the multiple objective optimization problem of Objective Function (4) in Section 2.2.3 well. The figures show that the remaining information is sufficient for high precision fault detection after executing conditional entropy-based feature selection.

Fig. 11 compares the accuracy of the equidistant feature selection [16] method and conditional entropy-based method. The equidistant feature selection in [16] is used to execute feature selection without measuring each feature. It is easy to implement. However, it may cause information loss. Fig. 11 shows that conditional entropy-based feature selection achieves lower misclassification rates (hence better accuracy). This is because conditional entropy-based feature selection measures each feature with the entropy.

4. CONCLUSION

In this study, a fault detection method with conditional entropy-based feature selection was presented for detecting parametric faults in analog circuits. The method involved two phases: the training phase and the diagnosis phase. In the two phases, feature selection was essential step to reduce computation overhead.

We observed the performance of three feature selection methods: equidistant method, conditional probability-based method and conditional entropy-based method. The equidistant method executed feature selection without measuring information loss. The conditional probability-based method used conditional probability to select features for minimizing information loss. However, feature selection in the conditional probability-based method can be changed into a multiple objective optimization problem. The study made the multiple objective optimization problem solvable in the conditional entropy-based method. In Section 3 we carried out three different feature selection methods, reducing the

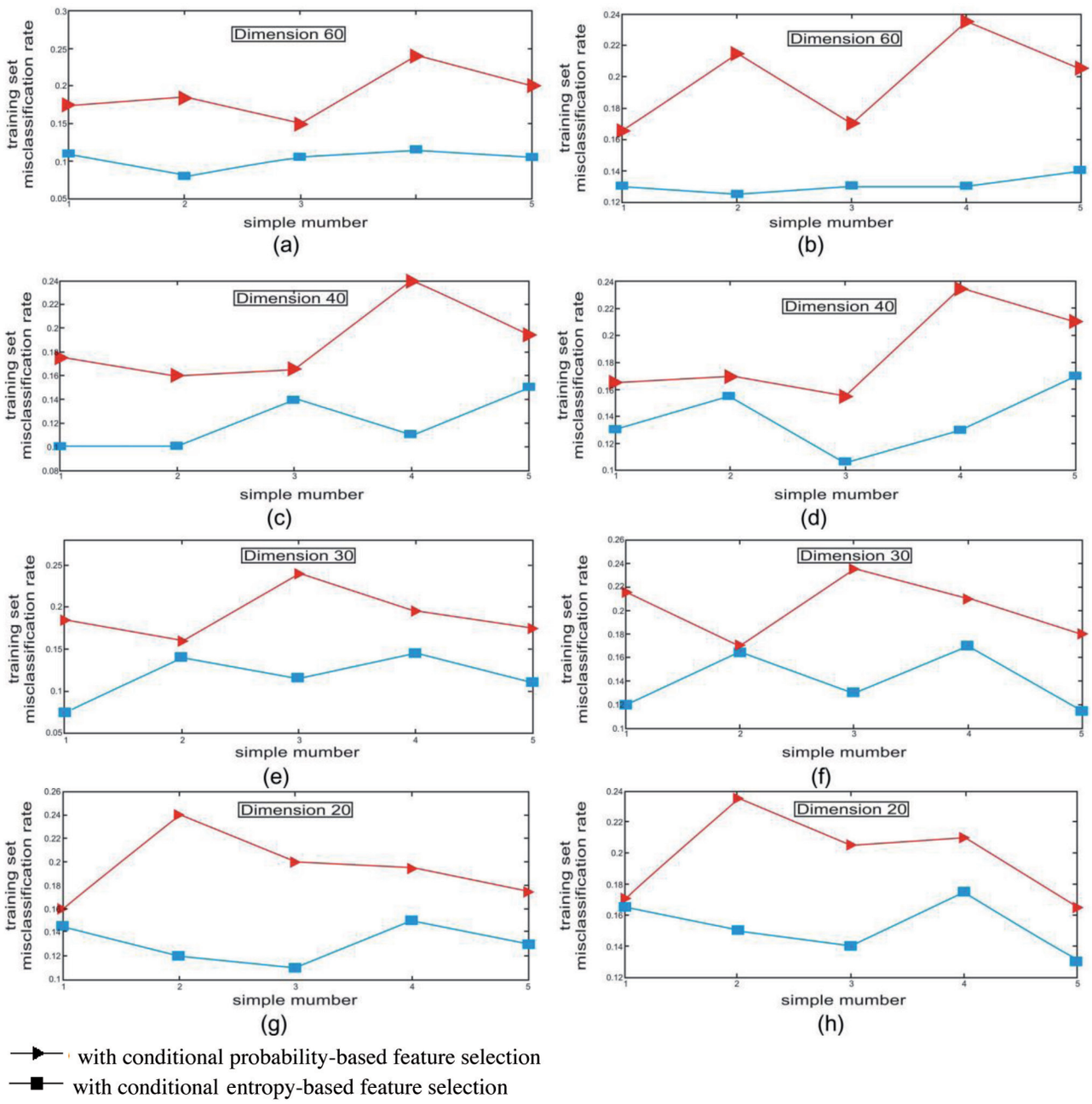


Fig. 9: Misclassification rate of fault detection with conditional probability-based and conditional entropy-based for three-pole active filter CUT. (a) training set of feature dimension 60. (b) testing set of feature dimension 60. (c) training set of feature dimension 40. (d) testing set of feature dimension 40. (e) training set of feature dimension 30. (f) testing set of feature dimension 30. (g) training set of feature dimension 20. (h) testing set of feature dimension 20

feature dimension from 120 to 60, 40, 30 and 20. In the practical application of the three methods in this study, the proposed conditional entropy-based feature selection reduced the feature dimension while it enhanced the fault detection accuracy. It better evaluated the information content of each raw observation vector dimension when selecting features in fault detection. In this study, the fault detection for parametric faults in analog circuits achieved low misclassification rates. The lowest misclassification rate of training sets for different samples of Fig. 7(a), Fig. 7(b) and Fig. 7(c) is 0.055%, 0.12% and 0.17%. The lowest misclassification rate of testing sets for different samples of Fig. 7(a), Fig. 7(b) and Fig. 7(c) is 0.135%, 0.13% and 0.19%. It showed that the SVM can separate non-linear space with optimal separating hyperplane.

BIBLIOGRAPHY

[1] Linda S, Milor. "A tutorial introduction to research on analog and mixed signal circuit testing". *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing*. October 1998. Vol. 45-10. p. 1389-1407. DOI: http://dx.doi.org/10.1109/82.728852

[2] Xiao Yingqun, Feng Lianggui. "A novel linear ridgelet network approach for analog fault diagnosis using wavelet-based fractal analysis and kernel PCA as preprocessors". *Measurement*. April 2012. Vol. 45-3. p. 297-310. DOI: http://dx.doi.org/10.1016/j.measurement.2011.11.018

[3] ChengLin Yang, Jing Yang, Zhen Liu, et al. "Complex Field Fault Modeling-Based Optimal Frequency Selection in Linear Analog Circuit Fault Diagnosis". *IEEE Transactions on Instrumentation and Measurement*. April 2014. Vol. 63-4. p. 813-825. DOI: http://dx.doi.org/10.1109/TIM.2013.2289074

[4] Vasan, Arvind Sai Sarathi, Long Bing, Pecht Michael. "Diagnostics and prognostics method for analog electronic circuits". *IEEE Transactions on Industrial Electronics*. October 2012. Vol.60-11. p. 5277-5291. DOI: http://dx.doi.org/10.1109/TIE.2012.2224074

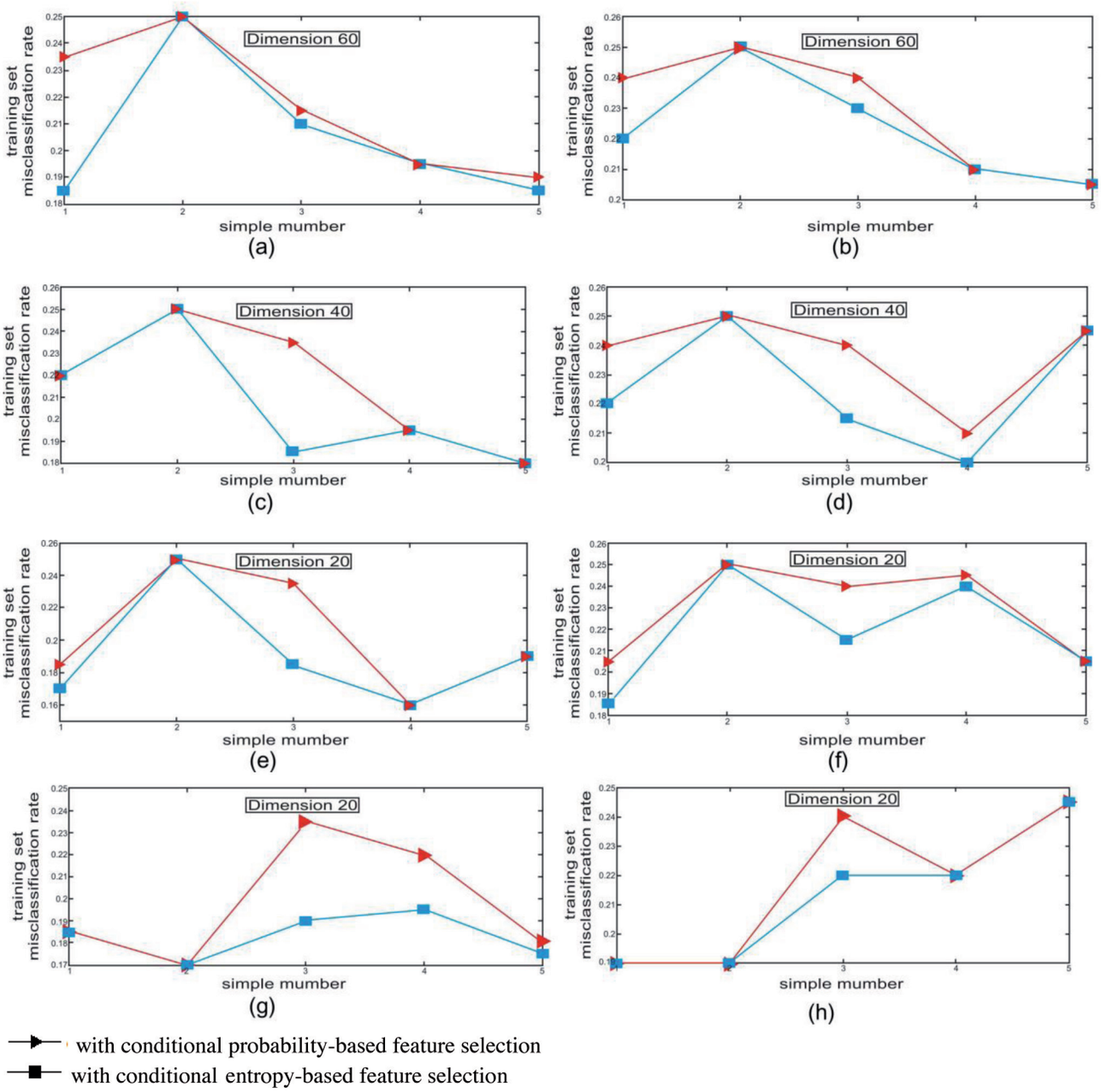


Fig. 10: Misclassification rate of fault detection with conditional probability-based and conditional entropy-based for five-pole active filter CUT. (a) training set of feature dimension 60. (b) testing set of feature dimension 60. (c) training set of feature dimension 40. (d) testing set of feature dimension 40. (e) training set of feature dimension 30. (f) testing set of feature dimension 30. (g) training set of feature dimension 20. (h) testing set of feature dimension 20

[5] Li Min, Xian Weiming, Long Bing, et al. "Prognostics of analog filters based on particle filters using frequency features". *Journal of Electronic Testing: Theory and Applications*. August 2013. Vol. 29-4. p. 567-584. DOI: http://dx.doi.org/10.1007/s10836-013-5383-y

[6] Ye Fangming, Zhang Zhaobo, Chakrabarty Krishnendu, et al. "Information-theoretic syndrome evaluation, statistical root-cause analysis, and correlation-based feature selection for guiding board-level fault diagnosis". *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*. June 2015. Vol.34-6. p.1014-1026. DOI: http://dx.doi.org/10.1109/TCAD.2015.2399438

[7] BenAli Jaouher, Saidi Lotfi, Mouelhi Aymen, et al. "Linear feature selection and classification using PNN and SFAM neural networks for a nearly online diagnosis of bearing naturally progressing degradations". *Engineering Applications of Artificial Intelligence*. June 2015. Vol.42. p. 67-81. DOI: http://dx.doi.org/10.1016/j.engappai.2015.03.013

[8] Li Zhihua. "A Novel Fault Diagnostic Method Based on Node-Voltage Vector Ambiguity Sets". *IEEE Transactions on Instrumentation and Measurement*. August 2014. Vol.63-8. p.1957-1965. DOI: http://dx.doi.org/10.1109/TIM.2014.2302236

[9] Shulin Tian, ChengLin Yang, Fang Chen, et al. "Circle Equation-Based Fault Modeling Method for Linear Analog Circuits". *IEEE Transactions on Instrumentation and Measurement*. September 2014. Vol.63-9. p.2145-2159. DOI: http://dx.doi.org/10.1109/TIM.2014.2307993

[10] Starzyk Janusz A, Liu Dong, Liu ZhiHong, et al. "Entropy-based optimum test points selection for analog fault dictionary techniques". *IEEE Transactions on Instrumentation and Measurement*. June 2004. Vol.53-3. p.754-761, DOI: http://dx.doi.org/10.1109/TIM.2004.827085

[11] Zhang Zhen, Duan Zhemin, Long Ying, et al. "A new swarm-SVM-based fault diagnosis approach for switched current circuit by using kurtosis and entropy as a preprocessor". *Analog Integrated Circuits and Signal Processing*. September 2014. Vol.81-1.p. 289-297. DOI: http://dx.doi.org/10.1007/s10470-014-0373-2

[12] Sun Jian, Wang Chenghua, Sun Jing, et al. "Analog circuit soft fault diagnosis based on PCA and PSO-SVM". *Journal of Networks*. December 2013. Vol. 8-12.p. 2791-2796. DOI: http://dx.doi.org/10.4304/jnw.8.12.2791-2796

[13] Achmad Widodo, Bo-Suk Yang. "Application of nonlinear feature extraction and support vector machines for fault diagnosis of induction motors". *Expert Systems with Applications*. July 2007. Vol.33-1. p.241-250. DOI: http://dx.doi.org/10.1016/j.eswa.2006.04.020

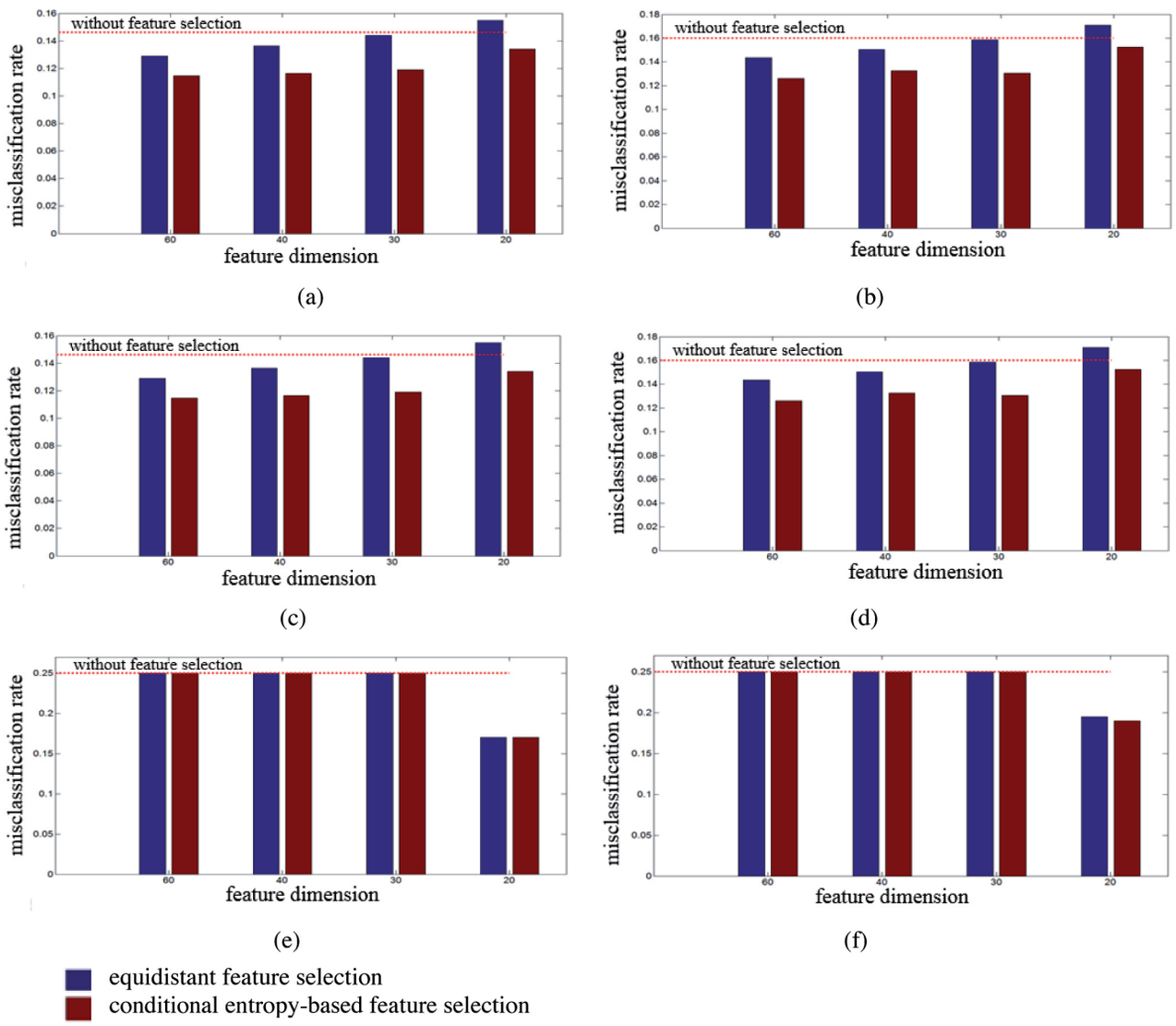


Fig. 11: Performance of conditional entropy-based feature selection. (a) misclassification rate of training set for two-pole active filter CUT. (b) misclassification rate of testing set for two-pole active filter CUT. (c) misclassification rate of training set for three-pole active filter CUT. (d) misclassification rate of testing set for three-pole active filter CUT. (e)misclassification rate of training set for five-pole active filter CUT. (f) misclassification rate of testing set for five-pole active filter CUT

[14] Wang Cong, Hefei AnHui, Gan Meng, et al. "Non-negative EMD manifold for feature extraction in machinery fault diagnosis". *Measurement: Journal of the International Measurement Confederation*. June 2015. Vol.70. p. 188-202. DOI: http://dx.doi.org/10.1016/j.measurement.2015.04.006

[15] Zhang Ying, Zuo Hongfu, Bai Fang. "Feature extraction for rolling bearing fault diagnosis by electrostatic monitoring sensors". *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science*. January 2015. Vol.229-10.p. 1887-1903. DOI: http://dx.doi.org/10.1177/0954406214550014

[16] ChenYang Pan, KwangTing Cheng. "Test generation for linear time-invariant analog circuits". *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing*. May 1999. Vol.46-5.p.554-564. DOI: http://dx.doi.org/10.1109/82.769804

[17] Liu Chao, Jiang Dongxiang, Yang Wenguang. "Global geometric similarity scheme for feature selection in fault diagnosis". *Expert Systems with Applications*. June 2014. Vol.41-8.p. 3585-3595. DOI: http://dx.doi.org/10.1016/j.eswa.2013.11.037

[18] Cerrada Mariela, Sánchez René Vinicio, Cabrera Diego et al. "Multi-stage feature selection by using genetic algorithms for fault diagnosis in gearboxes based on vibration signal". *Sensors (Switzerland)*. September 2015. Vol.15-9.p. 23903-23926. DOI:http://dx.doi.org/10.3390/s150923903

[19] Zhang Zhan'an, Cai Xingguo, Parkun Destung. "Determination of pumped storage capacity combining the entropy weighting method and principal component analysis". *Journal of Power Technologies*. 2014. Vol. 94-3.p.165-175

[20] Gray Robert M, Linder Tamás, Li Jia. "A Lagrangian formulation of Zador's entropy-constrained quantization theorem". *IEEE Transactions on Information Theory*. March 2002. Vol. 48-3. p. 695-707. DOI: http://dx.doi.org/10.1109/18.986007

[21] Keskes Hassen, Braham Ahmed. "Recursive Undecimated Wavelet Packet Transform and DAG SVM for Induction Motor Diagnosis". *IEEE Transactions on Industrial Informatics*. October 2015. Vol.11-5. p.1059-1066. DOI: http://dx.doi.org/10.1109/TII.2015.2462315

[22] Weiwu Yan, Huihe Shao. "Application of support vector machine nonlinear classifier to fault diagnoses". *Proceedings of the 4th World Congress on Intelligent Control and Automation*. June 2002. Vol.4. p.2697-2700. DOI: http://dx.doi.org/10.1109/WCICA.2002.1020004.

APPRECIATION

This research was funded by the Scientific Research Foundation of CUIT No. KYTZ201521.