Information Sciences 259 (2014) 252–268

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Bayesian classifiers based on probability density estimation and their applications to simultaneous fault diagnosis

Yu-Lin He [a,*], Ran Wang [b], Sam Kwong [b], Xi-Zhao Wang [a]

[a] College of Mathematics and Computer Science, Hebei University, Baoding 071002, Hebei, China
[b] Department of Computer Science, City University of Hong Kong, 83 Tat Chee Avenue, Kowloon, Hong Kong

ARTICLE INFO

Article history:
Available online 13 September 2013

Keywords:
Bayesian classification
Dependent feature
Joint probability density estimation
Optimal bandwidth
Simultaneous fault diagnosis
Single fault

ABSTRACT

A key characteristic of simultaneous fault diagnosis is that the features extracted from the original patterns are strongly dependent. This paper proposes a new model of Bayesian classifier, which removes the fundamental assumption of naive Bayesian, i.e., the independence among features. In our model, the optimal bandwidth selection is applied to estimate the class-conditional probability density function (p.d.f.), which is the essential part of joint p.d.f. estimation. Three well-known indices, i.e., classification accuracy, area under ROC curve, and probability mean square error, are used to measure the performance of our model in simultaneous fault diagnosis. Simulations show that our model is significantly superior to the traditional ones when the dependence exists among features.

1. Introduction

Fault diagnosis is the problem of detecting the potential faults hidden in the observed instances that are related to specific application domains. There are two types of fault diagnosis, i.e., single and simultaneous. In single fault diagnosis, only one fault may appear in an observed instance, while in simultaneous fault diagnosis, multiple faults may appear in an observed instance. Single fault diagnosis has been well studied in the past decade and has been applied to various domains, such as generator winding protection [1], chemical process [18], electrical machine [16], active magnetic bearing [20], power transformer [28], and field air defense gun [2]. Currently, with the development of science/technology, there is a stronger need on the safety and reliability of modern equipments. Unlike the traditional single fault generation, different faults often occur simultaneously in modern equipments due to various factors. Consequentially, these faults may cause serious accidents (e.g., air disasters, marine disasters, explosion accidents, collapse accidents, leakage accidents, and so on) that not only lead to great economic losses but also heavy casualties and environmental pollution. Therefore, an effective methodology is required to recognize the potential simultaneous faults in order to avoid such accidents. However, it is very difficult to conduct simultaneous fault diagnosis accurately and effectively due to the complex combination, mixture, and disturbance of features that reflect the single faults. A comprehensive reference-search finds that only a few literatures [4,10,24,27,29] exist to tackle this problem. These methods usually use the qualitative causal or quantitative analytical models to identify the simultaneous faults. Although a good solution is provided, these models usually cannot work well in practical applications. Meanwhile, the model parameters are also hard to determine.

* Corresponding author. Tel./fax: +86 312 5079638.
E-mail addresses: csylhe@gmail.com, yulinhe@ieee.org (Y.-L. He), ranwang3-c@my.cityu.edu.hk (R. Wang), cssamk@cityu.edu.hk (S. Kwong), xizhaowang@ieee.org (X.-Z. Wang).

0020-0255/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ins.2013.09.003

The main methodologies for handling simultaneous fault diagnosis include artificial neural networks (ANNs) [4,24], support vector machines (SVMs) [27,29], and Dempster–Shafer theory (DST) [10]. Different models have been designed for specific problems, i.e., chemical reactor [4,24], chemical plant [29], and multi-function rotor [10]. However, there are two disadvantages of the existing models. (1) The computational complexity for learning their parameters are high. Given that N is the size of training set, the training complexities of ANNs, SVMs, and DST are $O(N^2)$, $O(N^3)$, and $O(N^2)$ respectively, which make them unable to deal with large data. (2) They often neglect the necessary dependence among features in the observed instances, which exist in most practical applications. For example, in heart-disease electrocardiogram (ECG) [22], there is strong dependence between indices of vulnerability and heart rate [13]. These limitations motivate our idea in this paper to develop a novel simultaneous fault diagnosis model that can avoid the intractable complexity and take the dependence among features into account.

Naive Bayesian classifier (NBC) is a competent tool to deal with large data due to its simplicity, low computational complexity, and less memory requirement [3]. Applying NBC to fault diagnosis is an emergent research topic. Related studies on single fault diagnosis can be found from recent references [9,12,15,17]. To our best knowledge, we are the first one who try to establish a simultaneous fault diagnosis model based on Bayesian classifiers. In order to deal with the dependence among features, a non-naive Bayesian classifier (NNBC) is proposed to diagnose the possible faults hidden in the observed instances. It establishes a model of joint p.d.f. that is estimated by using Parzen windows based on the multivariate kernel function. Specifically, the estimation is completed by seeking an optimal bandwidth for the Parzen window through minimizing the mean integrated squared error between the true p.d.f. and the estimated p.d.f.

Analysis reveals that the training complexity of NNBC is $O(Nd)$, where N is the number of training instances and d is the number of conditional features. It shows that when $d \ll N$, NNBC can carry out the fault diagnosis with lower computational burden than ANNs [24], SVMs [29], and DST [10]. We compare our proposed NNBC with three p.d.f. density estimation based NBCs (normal naive Bayesian (NNB) [14], flexible naive Bayesian (FNB) [7], and the homologous model of FNB (FNB_{ROT})) [11] in terms of three evaluation indices, i.e., classification accuracy, area under ROC curve (AUC) [5,6], and probability mean square error (PMSE) [8]. The comparative results show that NNBC is uniformly and significantly superior to the other three models regarding the three indices, and therefore, provides a new way to design high-performance models for simultaneous fault diagnosis.

The rest of the paper is organized as follows: In Section 2, we summarize the basic naive Bayesian classifier algorithm. In Section 3, a non-naive Bayesian classification model based on the joint probability density estimation is proposed. In Section 4, we apply our proposed NNBC to simultaneous fault diagnosis. Finally, in Section 5, we conclude this paper and outline the main directions for future research.

2. A brief review on Bayesian classifiers

This section will give a brief review on naive Bayesian classifiers. We first introduce a number of denotations.

Let X be a set of N instances. Each instance is described by d condition attributes and one decision attribute. All the condition attributes are assumed to be continuous, and the decision attribute is supposed to be discrete. Suppose that the decision attribute takes values from $\{w_1, w_2, \ldots, w_c\}$, which implies that all instances are categorized into c classes. In this way, any instance in X will be denoted as a d-dimensional vector:

$$\tilde{x}_i^{(k)} = \left\{x_{i1}^{(k)}, x_{i2}^{(k)}, \ldots, x_{id}^{(k)}\right\} (1 \leqslant i \leqslant n_k, 1 \leqslant k \leqslant c),$$

where c is the number of classes and n_k is the number of instances within the kth class. Let $\tilde{x} = (x_1, x_2, \ldots, x_d)$ indicate a new example whose value of decision attribute is unknown.

Bayesian classifier [7,11,14,25,30] can assign the most likely class to the new example $\tilde{x} = (x_1, x_2, \ldots, x_d)$ by the Bayesian theorem. According to the prior probability and class conditional probability of the new example, Bayesian classifier calculates the posterior probability and determines the value of decision attribute for the new example. The Bayesian classifier discriminates the class of the new sample $\tilde{x}$ as the following equation:

$$w = \arg\max_{w_k, k=1,2,\ldots,c} \{P(w_k|\tilde{x})\} = \arg\max_{w_k, k=1,2,\ldots,c} \left\{\frac{P(w_k)P(\tilde{x}|w_k)}{P(\tilde{x})}\right\} = \arg\max_{w_k, k=1,2,\ldots,c} \{P(w_k)P(\tilde{x}|w_k)\}, \quad (1)$$

where $P(w_k)$ is the prior probability of the kth class, which can be estimated by the frequency of instances of the kth class, i.e., $P(w_k) \approx \frac{n_k}{N}$ in which $N = \sum_{k=1}^{c} n_k$ is the size of dataset X. $P(\tilde{x}|w_k)$ is called the class conditional probability. The crucial work of NBC is to estimate $P(\tilde{x}|w_k)$ based on the training instances in the kth class.

A fundamental assumption of the NBC is that all condition attributes are independent. Based on this assumption, the class conditional probability can be expressed as Eq. (2):

$$P(\tilde{x}|w_k) = P(x_1, x_k, \ldots, x_d|w_k) = \prod_{j=1}^{d} P(x_j|w_k). \quad (2)$$

By replacing the class conditional probability with Eq. (2), NBC can get the following decision rule (in Eq. (3)) for determining the value of decision attribute of $\tilde{x}$:

$$w = \arg\max_{w_k, k=1,2,\ldots,c} \left\{ \frac{n_k}{N} \prod_{j=1}^{d} P(x_j|w_k) \right\}. \tag{3}$$

From Eq. (3), we can find that the calculation of $P(x_j|w_k)(1 \leqslant j \leqslant d)$ is the key to classify new instance. There are three handling-methodologies based on the density estimation strategy to estimate the component $P(x_j|w_k)$ for $\tilde{x}$, i.e., NNB [14], FNB [7], and FNB_{ROT} [11].

2.1. NNB

NNB [14] assumes that the n_k values of the jth condition attribute, i.e., $x_{1j}^{(k)}, x_{2j}^{(k)}, \ldots, x_{n_k j}^{(k)}$, are generated from a single Gaussian distribution. Then, $P(x_j|w_k)$ can be calculated by Eq. (4):

$$P(x_j|w_k) = \frac{1}{\sqrt{2\pi}\sigma_j^{(k)}} \exp\left[-\frac{\left(x_j - \mu_j^{(k)}\right)^2}{2\left(\sigma_j^{(k)}\right)^2} \right], \tag{4}$$

where $\mu_j^{(k)} = \frac{\sum_{i=1}^{n_k} x_{ij}^{(k)}}{n_k}$ and $\left(\sigma_j^{(k)}\right)^2 = \frac{\sum_{i=1}^{n_k} \left[x_{ij}^{(k)} - \mu_j^{(k)}\right]^2}{n_k}$ are the mean value and variance of $x_{1j}^{(k)}, x_{2j}^{(k)}, \ldots, x_{n_k j}^{(k)}$ respectively.

NNB provides the simplest and fast way to estimate class conditional probability. The two required parameters, i.e., mean value and variance of the normal distribution, can be directly computed based on the given dataset. No sophisticated estimation strategy is needed.

2.2. FNB

The continuous attributes do not always follow the Gaussian distribution in many application domains. To cope with the case of non-Gaussian distribution, John and Langley [7] proposed the FNB which estimates $P(x_j|w_k)$ through the following Eq. (5):

$$P(x_j|w_k) = \frac{1}{n_k h_j^{(k)}} \sum_{i=1}^{n_k} \left[K\left(\frac{x_j - x_{ij}^{(k)}}{h_j^{(k)}} \right) \right], \tag{5}$$

where $h_j^{(k)}$ is the bandwidth and $K(*)$ is the kernel function. In FNB, $h_j^{(k)} = \frac{1}{\sqrt{n_k}}$ and $K(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$. This kernel is called Gaussian kernel. The experimental study shows that the classification performance of FNB mainly depends on the selection of the bandwidth $h_j^{(k)}$.

When the real distribution of the observed dataset is not held for normal, FNB can obtain a more accurate p.d.f. estimation compared with NNB due to the application of flexible density estimation method, i.e., Parzen window. Besides, the important bandwidth parameter is pre-assigned, which does not lead to the additional increase of training time.

2.3. FNB_{ROT}

In order to validate the impact of different parameter-selection methods on the classification performance, Liu et al. applied the rule of thumb [11] to the selection of bandwidth parameter of FNB. They replaced the traditional bandwidth parameter in FNB $h_j^{(k)} = \frac{1}{\sqrt{n_k}}$ with the following Eq. (6):

$$h_j^{(k)} = \left(\frac{4}{3n_k} \right)^{\frac{1}{5}} \sigma_j^{(k)}, \tag{6}$$

where $\left(\sigma_j^{(k)}\right)^2$ is the variance that can be calculated from the given instances $x_{1j}^{(k)}, x_{2j}^{(k)}, \ldots, x_{n_k j}^{(k)}$. In our study, we call this kind of Bayesian classifier FNB_{ROT}. In addition to the above-mentioned rule of thumb, we can find other methods of parameter selection from references (e.g. [19,23]). However, as the demonstrations in [11], the very sophisticated bandwidth selection schemes may not give good performance in the context of NBC classification, while some very simple schemes may give significantly better performance. Furthermore in [11], the simple scheme, i.e., rule of thumb, is used for bandwidth selection in their experiments.

The only difference between FNB and FNB_{ROT} is the determination of bandwidth parameter. FNB_{ROT} uses the rule of thumb scheme to estimate the bandwidth parameter, which guarantees a more appropriate bandwidth than FNB. The bandwidth in

FNB is totally independent from the given dataset, while the rule of thumb in FNB_{ROT} uses the information provided by the current dataset.

3. Bayesian model based on joint probability density estimation

As it is mentioned in Section 2, the fundamental assumption in NBC is that all conditional attributes are independent. In this section, we will propose an improved Bayesian classification model based on joint p.d.f. estimation, i.e., non-naive Bayesian classifier (NNBC), which releases the assumption of attribute-independence. First, the basic concept of joint p.d.f. estimation is introduced. Then, the optimal parameter selection in the joint p.d.f. estimation is discussed. Finally, the NNBC model is described in detail.

3.1. Joint p.d.f. estimation

In probability theory and statistical inference, p.d.f. estimation [19,23] refers to giving a specific function without unknown parameters such that the error between the function and the unobservable underlying p.d.f. can be small enough. Particularly, the estimation of p.d.f. for a continuous distribution from the representative samples is considered as one of the major ingredients in machine learning and pattern recognition. The well-known Parzen window method provides a consistent and asymptotic mode to approximately construct the underlying p.d.f.. Based on the set of d-dimensional data $\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_N$ where $\tilde{x}_i = (x_{i1}, x_{i2}, \ldots, x_{id})(1 \leqslant i \leqslant N)$, Parzen window method estimates the underlying joint p.d.f. through the following Eq. (7):

$$\hat{f}_h(\tilde{x}) = \frac{1}{Nh^d}\sum_{i=1}^{N}\left[K\left(\frac{\tilde{x}-\tilde{x}_i}{h}\right)\right] = \frac{1}{Nh^d}\sum_{i=1}^{N}\left[K\left(\frac{x_1-x_{i1}}{h}, \frac{x_2-x_{i2}}{h}, \ldots, \frac{x_d-x_{id}}{h}\right)\right], \tag{7}$$

where $K(*)$ is a multivariate kernel function and h is an important parameter, called bandwidth. The most common kernel is the multivariate Gaussian kernel as shown in Eq. (8):

$$K(\tilde{x}) = \frac{1}{\left(\sqrt{2\pi}\right)^d}\exp\left(-\frac{\tilde{x}\tilde{x}^T}{2}\right), \tag{8}$$

where $\tilde{x}^T$ is the transposition of vector $\tilde{x}$.

It is well acknowledged that the estimation performance of Parzen window method depends strongly on the selection of bandwidth h [11,19,23], which is related to the size of dataset N and should hold for the following two conditions:

$$\lim_{N\to+\infty} h(N) = 0 \text{ and } \lim_{N\to+\infty} N \times h(N) = +\infty.$$

Many researchers [19,23] have claimed that the appropriate selection of bandwidth can make the estimated error between the true p.d.f. and estimated p.d.f. converge or attain the minimum.

3.2. The optimal selection of bandwidth

In order to find the optimal bandwidth for joint p.d.f. estimation, in this section, we use the mean integrated squared error (MISE) to measure the difference between the true p.d.f. and the estimated p.d.f.. Let $f(\tilde{x})$ be the true p.d.f. of the observed data $\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_N$, then, MISE can be expressed as Eq. (9), in which $\int$ and $d\tilde{x}$ are the abbreviations of $\iint \ldots \int$ and $dx_1\, dx_2 \ldots dx_d$ respectively:

$$\mathrm{MISE}(h) = E\left[\int \{\hat{f}_h(\tilde{x}) - f(\tilde{x})\}^2 d\tilde{x}\right] = \int var\left(\hat{f}_h(\tilde{x})\right)d\tilde{x} + \int bias^2\left(\hat{f}_h(\tilde{x})\right)d\tilde{x}. \tag{9}$$

Then, we can derive the expressions of $bias\left(\hat{f}_h(\tilde{x})\right)$ and $var(\hat{f}_h(\tilde{x}))$ as follows:

$$\begin{aligned} bias\left(\hat{f}_h(\tilde{x})\right) &= E\left[\hat{f}_h(\tilde{x})\right] - f(\tilde{x}) = \int\left[\frac{1}{h^d}K\left(\frac{\tilde{x}-\tilde{y}}{h}\right)f(\tilde{y})\right]d\tilde{y} - f(\tilde{x}) = \int [K(\tilde{z})f(\tilde{x}-h\tilde{z})]d\tilde{z} - f(\tilde{x}) \\ &= \int\left\{K(\tilde{z})\left[f(\tilde{x}) - h\tilde{z}f'(\tilde{x}) + \frac{1}{2}h^2\tilde{z}\tilde{z}^T f''(\tilde{x}) + O(h^2) - f(\tilde{x})\right]\right\}d\tilde{z} \\ &= -hf'(\tilde{x})\int \tilde{z}K(\tilde{z})d\tilde{z} + \frac{1}{2}h^2 f''(\tilde{x})\int \tilde{z}\tilde{z}^T K(\tilde{z})d\tilde{z} + O(h^2)\int f(\tilde{z})d\tilde{z}, \end{aligned} \tag{10}$$

and

$$\begin{aligned} var\left(\hat{f}_h(\tilde{x})\right) &= E\left\{\hat{f}_h(\tilde{x}) - E\left[\hat{f}_h(\tilde{x})\right]\right\}^2 = E\left[\hat{f}_h(\tilde{x})\right]^2 - \left\{E\left[\hat{f}_h(\tilde{x})\right]\right\}^2 \\ &= \frac{1}{N}\int\left[\frac{1}{h^{2d}}K\left(\frac{\tilde{x}-\tilde{y}}{h}\right)^2 f(\tilde{y})\right]d\tilde{y} - \frac{1}{N}\left\{\int\left[\frac{1}{h^{2d}}K\left(\frac{\tilde{x}-\tilde{y}}{h}\right)f(\tilde{y})\right]d\tilde{y}\right\}^2 \\ &= \frac{1}{Nh^d}\int[K(\tilde{z})^2 f(\tilde{x}-h\tilde{z})]d\tilde{z} - \frac{1}{Nh^d}\left\{\int[K(\tilde{z})f(\tilde{x}-h\tilde{z})]d\tilde{z}\right\}^2 \\ &= \frac{1}{Nh^d}\left\{f(\tilde{x})\int K(\tilde{z})^2 d\tilde{z} - hf'(\tilde{x})\int \tilde{z}K(\tilde{z})^2 d\tilde{z} + \frac{1}{2}h^2 f''(\tilde{x})\int \tilde{z}\tilde{z}^T K(\tilde{z})^2 d\tilde{z} + O(h^2)\right\} + O(N^{-1}), \end{aligned} \tag{11}$$

where, $\tilde{z} = \frac{\tilde{x}-\tilde{y}}{h}$.

It is known that for the multivariate Gaussian kernel $K(\tilde{z})$, $\int \tilde{z}K(\tilde{z})d\tilde{z} = 0$ and $\int K(\tilde{z})d\tilde{z} = 1$ hold well. Substituting these 2 integrals in Eq. (10), we have the following Eq. (12):

$$bias\left(\hat{f}_h(\tilde{x})\right) = \frac{1}{2}h^2 f''(\tilde{x})\int \tilde{z}\tilde{z}^T K(\tilde{z})d\tilde{z} + O(h^2). \tag{12}$$

Having

$$\frac{1}{Nh^d}[hf'(\tilde{x})\int \tilde{z}K(\tilde{z})^2 d\tilde{z}] = O(N^{-1}),$$

and

$$\frac{1}{Nh^d}\left[\frac{1}{2}h^2 f''(\tilde{x})\int \tilde{z}\tilde{z}^T K(\tilde{z})^2 d\tilde{z}\right] = O(N^{-1}),$$

the expression of $var\left(\hat{f}_h(\tilde{x})\right)$ can be rewritten as:

$$var\left(\hat{f}_h(\tilde{x})\right) = \frac{1}{Nh^d}f(\tilde{x})\int K(\tilde{z})^2 d\tilde{z} + O(N^{-1}h^{-d}). \tag{13}$$

Neglecting the terms $O(h^2)$ and $O(N^{-1}h^{-d})$ in Eqs. (12) and (13) when $h \to 0$ and $Nh \to +\infty$, and replacing $bias(\hat{f}_h(\tilde{x}))$ and $var(\hat{f}_h(\tilde{x}))$ in Eq. (9) with the derived Eqs. (12) and (13) respectively, we can get the following Eq. (14):

$$\mathrm{MISE}(h) = \frac{1}{Nh^d}\left[\int K(\tilde{z})^2 d\tilde{z}\right]\left[\int f(\tilde{x})d\tilde{x}\right] + \frac{1}{4}h^4\left[\int \tilde{z}\tilde{z}^T K(\tilde{z})d\tilde{z}\right]^2\left\{\int [f''(\tilde{x})]^2 d\tilde{x}\right\}. \tag{14}$$

Let $R(K) = \int K(\tilde{z})^2 d\tilde{z}$, $\mu_2(K) = \int \tilde{z}\tilde{z}^T K(\tilde{z})d\tilde{z}$, and $R(f'') = \int [f''(\tilde{x})]^2 d\tilde{x}$. Note that $\int f(\tilde{x})d\tilde{x} = 1$, thus we can simplify the expression of MISE (h) as the following Eq. (15):

$$\mathrm{MISE}(h) = \frac{1}{Nh^d}[R(K)] + \frac{1}{4}h^4[\mu_2(K)]^2 R(f''). \tag{15}$$

To find the optimal bandwidth that can make MISE (h) reach the minimum, we let the first order partial derivative of MISE (h) with respect to h be 0, i.e., $\frac{d\mathrm{MISE}(h)}{dh} = 0$ which implies that the optimal h is attained at

$$h_{\text{optimal}}^{(\mathrm{MISE})} = \left[\frac{dR(K)}{[\mu_2(K)]^2 R(f'')N}\right]^{\frac{1}{d+4}}, \tag{16}$$

and the corresponding minimal MISE (h) is given by

$$\inf_{h>0}\mathrm{MISE}(h) = \frac{d+4}{4d}\left\{[\mu_2(K)]^{2d}[dR(K)]^4[R(f'')]^d N^{-4}\right\}^{\frac{1}{d+4}}. \tag{17}$$

In the following we point out how to compute the 3 components $R(K)$, $\mu_2(K)$, and $R(f'')$ in Eqs. (16) and (17). For the multivariate Gaussian kernel, we can calculate

$$R(K) = \frac{1}{\left(\sqrt{2\pi}\right)^{2d}}\prod_{j=1}^{d}\int \exp\left(-x_j^2\right)dx_j = (4\pi)^{-\frac{d}{2}}, \tag{18}$$

and

$$\mu_2(K) = \frac{1}{\left(\sqrt{2\pi}\right)^{d}}\sum_{j=1}^{d}\left[\int x_j^2 \exp\left(-\frac{x_j^2}{2}\right)dx_j\right] = 1. \tag{19}$$

For the sake of robust estimation, we consider $f(\tilde{x})$ as a multivariate normal density function $N(0, \Sigma)$ with the diagonal matrix $\Sigma = \mathrm{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_d^2)$ where $\sigma_j^2, (1 \leqslant j \leqslant d)$ is the variance of $x_{1j}, x_{2j}, \ldots, x_{Nj}$. We consider a special case of $d = 2$ and give its derivation of $R(f')$ where $f(\tilde{x}) = f(x_1, x_2) = \frac{1}{(\sqrt{2\pi})^2 \sigma_1 \sigma_2} \exp\left[-\left(\frac{x_1^2}{2\sigma_1^2} + \frac{x_2^2}{2\sigma_2^2}\right)\right]$. Then, the formula of $R[f'(x_1, x_2)]$ can be expressed as Eq. (20):

$$R[f''(x_1, x_2)] = \frac{1}{4\left(\sqrt{2\pi}\right)^2 \sigma_1 \sigma_2}\left[2\left(\frac{1}{\sigma_1^4} + \frac{1}{\sigma_2^4}\right) + \left(\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}\right)^2\right]. \tag{20}$$

Similarly, the derivation of $R[f''(\tilde{x})]$ for the p.d.f. estimation with d variables $(d > 2)$ can be given by Eq. (21):

$$R[f''(\tilde{x})] = \frac{2\sum_{j=1}^{d} \frac{1}{\sigma_j^4} + \left(\sum_{j=1}^{d} \frac{1}{\sigma_j^2}\right)^2}{4\left(\sqrt{2\pi}\right)^d \prod_{j=1}^{d} \sigma_j} \tag{21}$$

i.e.,

$$R(f'') = \frac{(4\pi)^{-\frac{d}{2}} |\Sigma|^{-\frac{1}{2}} \{2\mathrm{tr}(\Sigma^{-1}\Sigma^{-1}) + \mathrm{tr}^2(\Sigma^{-1})\}}{4}. \tag{22}$$

Bringing Eqs. 18, 19, 22 into Eqs. (16) and (17), we can get the optimal bandwidth in Eq. (23)

$$h_{\mathrm{optimal}}^{(\mathrm{MISE})} = \left(\frac{4d}{N|\Sigma|^{-\frac{1}{2}}\{2\mathrm{tr}(\Sigma^{-1}\Sigma^{-1}) + \mathrm{tr}^2(\Sigma^{-1})\}}\right)^{\frac{1}{d+4}} \tag{23}$$

and the minimal MISE in Eq. (24)

$$\inf_{h>0} \mathrm{MISE}(h) = (4\pi)^{-\frac{d}{2}}\left(\frac{d+4}{4d}\right)\left(\frac{d}{N}\right)^{\frac{4}{d+4}}\left(\frac{|\Sigma|^{-\frac{1}{2}}\{2\mathrm{tr}(\Sigma^{-1}\Sigma^{-1}) + \mathrm{tr}^2(\Sigma^{-1})\}}{4}\right)^{\frac{d}{d+4}}. \tag{24}$$

3.3. Non-naive Bayesian classifier

As discussed in the previous sections, NNB, FNB and $\mathrm{FNB}_{\mathrm{ROT}}$ have the following two restrictions. (1) They are based on such an assumption that each condition attribute is independent of any other one, which obviously does not hold in many real-world applications. (2) When estimating the marginal p.d.f., NNB assumes that each attribute follows a normal distribution, while FNB/$\mathrm{FNB}_{\mathrm{ROT}}$ do not have an appropriate strategy of parameter selection, which seriously affects the estimation precision. In order to relax the above-mentioned two restrictions, we propose the NNBC, which removes the independence among attributes, and replaces the marginal p.d.f. estimations by joint p.d.f. estimation. NNBC determines the class of a new sample $\tilde{x}$ as the following Eq. (25),

$$w = \arg\max_{w_k, k=1,2,\ldots,c}\left\{\frac{n_k}{N}P(\tilde{x}|w_k)\right\} = \arg\max_{w_k, k=1,2,\ldots,c}\left\{\frac{1}{Nh_k^d}\sum_{i=1}^{n_k}\left[K\left(\frac{x_1 - x_{i1}^{(k)}}{h_k}, \frac{x_2 - x_{i2}^{(k)}}{h_k}, \ldots, \frac{x_d - x_{id}^{(k)}}{h_k}\right)\right]\right\}, \tag{25}$$

where $K(\tilde{x}) = \frac{1}{(\sqrt{2\pi})^d} \exp\left(-\frac{\tilde{x}\tilde{x}^T}{2}\right)$ is the multivariate Gaussian kernel as shown in Eq. (8), $h_k (1 \leqslant k \leqslant c)$ is the optimal bandwidth which has been derived as in Section 3.2.

Specifically, for a set of instances belonging to the kth class, the optimal bandwidth $h_k (1 \leqslant k \leqslant c)$ given in Eq. (23) can be simplified as:

$$h_k = \left(\frac{4d}{n_k|\Sigma_k|^{-\frac{1}{2}}\left\{2\mathrm{tr}\left(\Sigma_k^{-1}\Sigma_k^{-1}\right) + \mathrm{tr}^2\left(\Sigma_k^{-1}\right)\right\}}\right)^{\frac{1}{d+4}},$$

where

$$\Sigma_k = \mathrm{diag}\left\{\left[\sigma_1^{(k)}\right]^2, \left[\sigma_2^{(k)}\right]^2, \ldots, \left[\sigma_d^{(k)}\right]^2\right\}.$$

The main differences among NNB, FNB, FNB_{ROT}, and our proposed NNBC are summarized as follows:

(1) NNB, FNB, and FNB_{ROT} assume that all condition attributes are independent. By calculating every component $P(x_j|w_k)(1 \leqslant j \leqslant d, 1 \leqslant k \leqslant c)$ through the marginal p.d.f., NNB, FNB, and FNB_{ROT} get the class conditional probability $P(\tilde{x}|w_k)$ for the new instance $\tilde{x} = \{x_1, x_2, \ldots, x_d\}$. Our proposed NNBC, which removes the independence assumption, establishes a model of joint p.d.f. in the estimation of $P(\tilde{x}|w_k)$ based on the multivariate kernel function.
(2) Due to the inappropriate distribution assumption in NNB and the non-optimal parameter selection in FNB and FNB_{ROT}, there are usually large errors between the true p.d.f. and the estimated p.d.f.. The imprecise estimation of p.d.f. for NBC will lead to the dissatisfactory classification performance. By minimizing the MISE, our proposed NNBC finds the optimal bandwidth for the joint p.d.f. estimation, which makes the estimated error reach the minimum.

Now, we give an analysis on the time complexity of the above-mentioned four Bayesian classification algorithms, i.e., NNB, FNB, FNB_{ROT}, and NNBC. Let N be the number of training instances, M the number of testing instances, and d the number of condition attributes. Since NNB needs to calculate the means and variances of the d condition attributes, the training and testing complexities are $O(Nd)$ and $O(Md)$ respectively. FNB uses the superposition of N p.d.f.s of the normal distribution to fit the true p.d.f., thus, the training and testing complexities are $O(Nd)$ and $O(MNd)$ respectively. In comparison with FNB, the application of the rule of thumb in FNB_{ROT} leads to some increase in the training time, but the training and testing complexities are still $O(Nd)$ and $O(MNd)$ respectively. Similar to FNB_{ROT}, our NNBC also costs extra training time to compute the optimal bandwidth. However, the parameter determination does not lead to additional increase of testing time. Thus, the training and testing complexities of NNBC are still $O(Nd)$ and $O(MNd)$ respectively.

4. Application to simultaneous faults diagnosis

In this section, we first design a device that can generate instances with single and simultaneous faults, then we demonstrate the performance of NNBC in simultaneous fault diagnosis on instances with strong dependence among features.

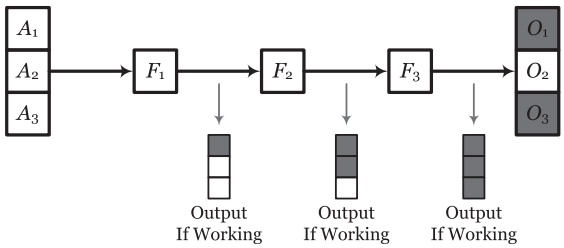


Fig. 1. The single and simultaneous faults generation.

Table 1
The detailed experimental results of classification accuracy and standard deviation on 20 single fault datasets.

	Datasets	NNB	FNB	FNB_{ROT}	NNBC
1	SinglFG (50)	0.721 ± 0.019	0.601 ± 0.038	0.703 ± 0.024	0.888 ± 0.008
2	SinglFG (100)	0.648 ± 0.007	0.590 ± 0.018	0.661 ± 0.011	0.890 ± 0.006
3	SinglFG (150)	0.703 ± 0.017	0.589 ± 0.019	0.682 ± 0.018	0.891 ± 0.012
4	SinglFG (200)	0.658 ± 0.014	0.614 ± 0.010	0.649 ± 0.010	0.896 ± 0.010
5	SinglFG (250)	0.691 ± 0.012	0.637 ± 0.013	0.702 ± 0.011	0.896 ± 0.010
6	SinglFG (300)	0.666 ± 0.007	0.590 ± 0.018	0.665 ± 0.009	0.927 ± 0.009
7	SinglFG (350)	0.698 ± 0.004	0.635 ± 0.011	0.703 ± 0.011	0.918 ± 0.007
8	SimulFG (400)	0.645 ± 0.005	0.600 ± 0.014	0.693 ± 0.011	0.896 ± 0.005
9	SinglFG (450)	0.709 ± 0.008	0.634 ± 0.014	0.692 ± 0.006	0.950 ± 0.008
10	SinglFG (500)	0.687 ± 0.009	0.642 ± 0.010	0.694 ± 0.007	0.922 ± 0.006
11	SinglFG (550)	0.682 ± 0.008	0.618 ± 0.018	0.679 ± 0.008	0.920 ± 0.006
12	SinglFG (600)	0.679 ± 0.006	0.637 ± 0.010	0.667 ± 0.006	0.934 ± 0.006
13	SinglFG (650)	0.699 ± 0.007	0.632 ± 0.015	0.680 ± 0.006	0.933 ± 0.005
14	SinglFG (700)	0.695 ± 0.007	0.631 ± 0.010	0.673 ± 0.007	0.920 ± 0.006
15	SinglFG (750)	0.740 ± 0.007	0.660 ± 0.009	0.723 ± 0.009	0.941 ± 0.009
16	SinglFG (800)	0.676 ± 0.005	0.650 ± 0.008	0.702 ± 0.005	0.949 ± 0.005
17	SinglFG (850)	0.701 ± 0.006	0.631 ± 0.007	0.684 ± 0.007	0.956 ± 0.006
18	SinglFG (900)	0.713 ± 0.004	0.657 ± 0.006	0.716 ± 0.005	0.941 ± 0.003
19	SinglFG (950)	0.685 ± 0.005	0.620 ± 0.008	0.683 ± 0.005	0.939 ± 0.006
20	SinglFG (1000)	0.676 ± 0.003	0.607 ± 0.010	0.678 ± 0.005	0.961 ± 0.004
Average		0.689 ± 0.008	0.624 ± 0.013	0.686 ± 0.009	0.923 ± 0.007

4.1. Single and simultaneous fault generation

Simultaneous fault diagnosis can be regarded as a multi-class classification problem in which every observed instance may be classified into more than two classes. In order to get the classifiers, a number of training instances associated with single or simultaneous faults are required. We develop a device as shown in Fig. 1 to generate these instances. In Fig. 1, A_i ($i = 1, 2, 3$) (for simplicity, we only consider the case of 3 input units in this paper.) is the ith input unit. F_i ($i = 1, 2, 3$) is the ith switch function that can transform the ith input A_i into $sin(A_i)$ or $cos(A_i)$. $sin(A_i)$ indicates that the switch function F_i is working and no fault occurs when the input passes through it. Otherwise, $cos(A_i)$ shows that the switch function F_i is not working and the corresponding fault occurs when the input passes through it. If all switch functions work effectively on the inputs, then the 3 outputs in Fig. 1 should be covered with black color. However, we find that the second output is not marked, and the switch function F_2 has no effect on the input that passes through it. In our study, the single or simultaneous fault instance I is depicted as the following vector:

Table 2
The detailed experimental results of ranking performance and standard deviation on 20 single fault datasets.

	Datasets	NNB	FNB	FNB_{ROT}	NNBC
1	SinglFG (50)	0.836 ± 0.015	0.746 ± 0.021	0.826 ± 0.018	0.988 ± 0.004
2	SinglFG (100)	0.758 ± 0.015	0.676 ± 0.009	0.728 ± 0.014	0.967 ± 0.003
3	SinglFG (150)	0.837 ± 0.006	0.757 ± 0.013	0.801 ± 0.008	0.983 ± 0.003
4	SinglFG (200)	0.774 ± 0.009	0.694 ± 0.014	0.769 ± 0.006	0.982 ± 0.003
5	SinglFG (250)	0.886 ± 0.003	0.820 ± 0.007	0.861 ± 0.005	0.988 ± 0.001
6	SinglFG (300)	0.829 ± 0.003	0.785 ± 0.011	0.825 ± 0.008	0.990 ± 0.002
7	SinglFG (350)	0.880 ± 0.003	0.799 ± 0.010	0.860 ± 0.006	0.991 ± 0.002
8	SinglFG (400)	0.855 ± 0.004	0.782 ± 0.005	0.834 ± 0.005	0.986 ± 0.002
9	SinglFG (450)	0.856 ± 0.003	0.808 ± 0.010	0.846 ± 0.005	0.992 ± 0.001
10	SinglFG (500)	0.842 ± 0.004	0.793 ± 0.005	0.834 ± 0.006	0.984 ± 0.003
11	SinglFG (550)	0.869 ± 0.003	0.818 ± 0.010	0.863 ± 0.004	0.991 ± 0.002
12	SinglFG (600)	0.818 ± 0.003	0.740 ± 0.006	0.795 ± 0.005	0.992 ± 0.001
13	SinglFG (650)	0.834 ± 0.002	0.781 ± 0.010	0.824 ± 0.004	0.993 ± 0.001
14	SinglFG (700)	0.860 ± 0.002	0.781 ± 0.008	0.836 ± 0.004	0.991 ± 0.002
15	SinglFG (750)	0.858 ± 0.004	0.795 ± 0.005	0.847 ± 0.003	0.992 ± 0.001
16	SinglFG (800)	0.879 ± 0.002	0.822 ± 0.004	0.867 ± 0.002	0.996 ± 0.001
17	SinglFG (850)	0.849 ± 0.002	0.792 ± 0.005	0.834 ± 0.002	0.995 ± 0.001
18	SinglFG (900)	0.860 ± 0.003	0.811 ± 0.005	0.844 ± 0.004	0.995 ± 0.002
19	SinglFG (950)	0.859 ± 0.002	0.802 ± 0.006	0.847 ± 0.002	0.994 ± 0.001
20	SinglFG (1000)	0.884 ± 0.002	0.829 ± 0.004	0.868 ± 0.003	0.997 ± 0.001
Average		0.846 ± 0.005	0.782 ± 0.008	0.830 ± 0.006	0.989 ± 0.002

Table 3
The detailed experimental results of estimation quality and standard deviation on 20 single fault datasets.

	Datasets	NNB	FNB	FNB_{ROT}	NNBC
1	SinglFG (50)	0.423 ± 0.010	0.526 ± 0.020	0.444 ± 0.014	0.172 ± 0.005
2	SinglFG (100)	0.403 ± 0.010	0.508 ± 0.016	0.426 ± 0.012	0.152 ± 0.007
3	SinglFG (150)	0.398 ± 0.009	0.503 ± 0.014	0.428 ± 0.007	0.174 ± 0.005
4	SinglFG (200)	0.443 ± 0.007	0.557 ± 0.015	0.459 ± 0.006	0.184 ± 0.006
5	SinglFG (250)	0.365 ± 0.005	0.522 ± 0.010	0.407 ± 0.006	0.130 ± 0.006
6	SinglFG (300)	0.376 ± 0.004	0.503 ± 0.011	0.424 ± 0.006	0.140 ± 0.005
7	SinglFG (350)	0.438 ± 0.006	0.552 ± 0.006	0.451 ± 0.005	0.142 ± 0.003
8	SinglFG (400)	0.420 ± 0.004	0.551 ± 0.008	0.450 ± 0.005	0.150 ± 0.002
9	SinglFG (450)	0.349 ± 0.002	0.426 ± 0.007	0.365 ± 0.003	0.117 ± 0.003
10	SinglFG (500)	0.435 ± 0.003	0.543 ± 0.007	0.465 ± 0.004	0.146 ± 0.002
11	SinglFG (550)	0.390 ± 0.003	0.488 ± 0.007	0.412 ± 0.004	0.110 ± 0.004
12	SinglFG (600)	0.418 ± 0.002	0.548 ± 0.008	0.445 ± 0.003	0.152 ± 0.002
13	SinglFG (650)	0.409 ± 0.003	0.504 ± 0.009	0.434 ± 0.004	0.124 ± 0.002
14	SinglFG (700)	0.408 ± 0.003	0.517 ± 0.007	0.434 ± 0.004	0.113 ± 0.002
15	SinglFG (750)	0.395 ± 0.003	0.496 ± 0.008	0.420 ± 0.003	0.135 ± 0.002
16	SinglFG (800)	0.396 ± 0.003	0.486 ± 0.010	0.410 ± 0.004	0.109 ± 0.002
17	SinglFG (850)	0.412 ± 0.002	0.513 ± 0.007	0.430 ± 0.002	0.128 ± 0.001
18	SinglFG (900)	0.378 ± 0.003	0.461 ± 0.005	0.394 ± 0.003	0.107 ± 0.002
19	SinglFG (950)	0.394 ± 0.002	0.482 ± 0.005	0.414 ± 0.001	0.109 ± 0.002
20	SinglFG (1000)	0.417 ± 0.001	0.512 ± 0.007	0.442 ± 0.002	0.118 ± 0.002
Average		0.403 ± 0.004	0.510 ± 0.009	0.428 ± 0.005	0.136 ± 0.003

$$I = \{A_1, A_2, A_3, O_1, O_2, O_3\},$$

where if $O_i = cos(A_i)$ $(i = 1, 2, 3)$, the ith fault exists in I; if $O_i = sin(A_i)$ $(i = 1, 2, 3)$, the ith fault does not occur in I. For example, $I' = \{A_1, A_2, A_3, sin(A_1), cos(A_2), sin(A_3)\}$ has fault in the second position and $I'' = \{A_1, A_2, A_3, cos(A_1), sin(A_2), cos(A_3)\}$ has faults in the first and third positions simultaneously.

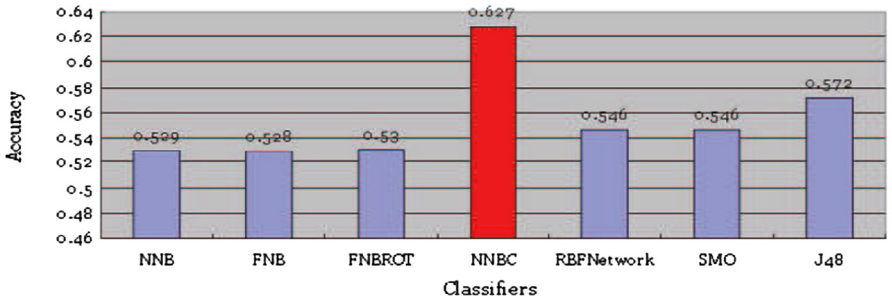
As mentioned above, our proposed NNBC removes the independence assumption in the traditional NBC and can deal with classification problem when features are strongly dependent. In order to guarantee that the features are dependent, we use the random vector conforming to the following probability distribution as the input of (A_1, A_2, A_3):

$$f(\tilde{x}) = \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\tilde{x}-\mu)\Sigma^{-1}(\tilde{x}-\mu)^T\right),$$

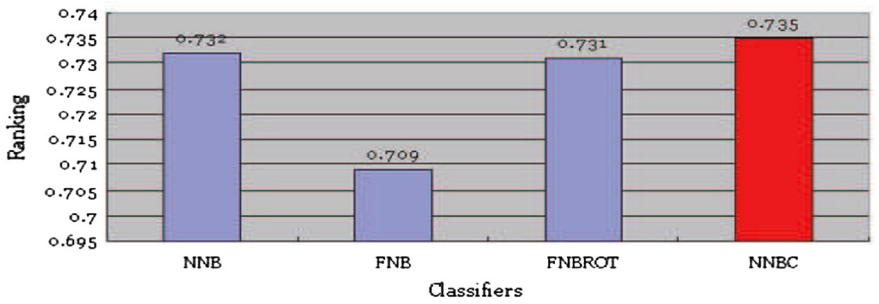
where $d = 3$, $\mu = [0, 0, 0]$, $\Sigma = [1, 0.75, 0.75; 0.75, 1, 0.75; 0.75, 0.75, 1]$, and $\tilde{x}^T$ is the transposition of vector $\tilde{x}$.

4.2. Simultaneous fault diagnosis based on proposed NNBC

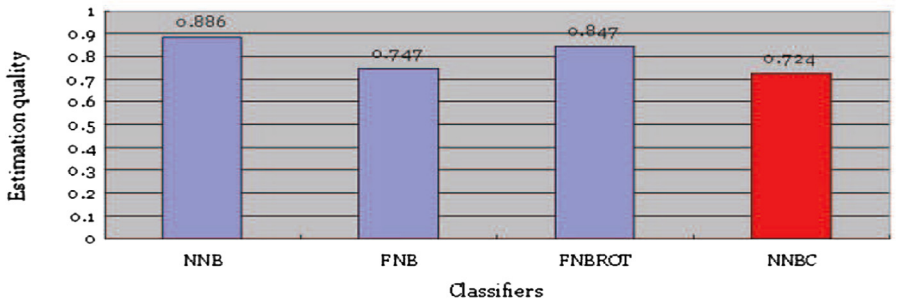
We first compare NNBC with NNB, FNB, and FNB_{ROT} in single fault diagnosis. Three indices, i.e., classification accuracy, AUC [5,6], and PMSE [8], are used to evaluate the different learning models. AUC is based on such a fact that the cost of classifying a sample into the wrong class is significantly lower than the reverse. A higher AUC value indicates that the



(a) The classification accuracies on Steel Plates Faults dataset



(b) The ranking performances on Steel Plates Faults dataset



(c) The estimated qualities on Steel Plates Faults dataset

Fig. 2. Experimental comparisons of different diagnosis methods on steel plates faults dataset.

corresponding learning model has a lower misclassification cost. PMSE is to measure the quality of class-conditional p.d.f. estimation. A lower PMSE represents a higher quality of estimation. The comparative results on 20 single fault datasets in terms of the three indices are summarized in Tables 1–3 respectively. In these tables, SinglFG (n) is the dataset in which all instances are generated with one single fault and every single fault contains n instances. Then, the size of dataset SinglFG (n) is $3n$. We conduct 10-fold cross-validation for 10 times. Each time, the same training and testing sets are used for the four methods, then the evaluations on classification accuracy, AUC, and PMSE are simultaneously performed.

From Tables 1–3, we can find that NNBC obtains the best classification accuracy, ranking performance, and estimation quality on these 20 single fault datasets. The performances of NNBC are far superior to those of NNB, FNB, and FNB_{ROT}. The average accuracy, AUC, and PMSE of NNBC are 0.923, 0.989, and 0.136 respectively, which indicate that it achieves the most accurate and effective diagnosis for single faults among the compared models.

We also test the four methods on a selected UCI dataset named steel plates faults [21]. This database has 1941 instances related to seven types of steel plates faults. Each instance has 25 continuous and 2 nominal attributes. In our experiments, we only consider the 25 continuous attributes. Meanwhile, we also compare the classification accuracy of Bayesian models

Table 4
The detailed experimental results of classification accuracy and standard deviation on 20 simultaneous fault datasets.

	Datasets	NNB	FNB	FNB_{ROT}	NNBC
1	SimulFG (50)	0.757 ± 0.013	0.689 ± 0.020	0.715 ± 0.015	0.883 ± 0.016
2	SimulFG (100)	0.674 ± 0.014	0.617 ± 0.011	0.674 ± 0.009	0.848 ± 0.007
3	SimulFG (150)	0.669 ± 0.007	0.656 ± 0.015	0.656 ± 0.010	0.871 ± 0.010
4	SimulFG (200)	0.700 ± 0.011	0.649 ± 0.011	0.668 ± 0.014	0.899 ± 0.006
5	SimulFG (250)	0.647 ± 0.007	0.636 ± 0.009	0.667 ± 0.010	0.899 ± 0.008
6	SimulFG (300)	0.685 ± 0.008	0.634 ± 0.012	0.693 ± 0.007	0.934 ± 0.008
7	SimulFG (350)	0.675 ± 0.009	0.628 ± 0.009	0.666 ± 0.006	0.903 ± 0.011
8	SimulFG (400)	0.681 ± 0.008	0.664 ± 0.007	0.686 ± 0.007	0.916 ± 0.006
9	SimulFG (450)	0.745 ± 0.008	0.694 ± 0.012	0.739 ± 0.008	0.937 ± 0.011
10	SimulFG (500)	0.700 ± 0.004	0.678 ± 0.007	0.706 ± 0.008	0.924 ± 0.004
11	SimulFG (550)	0.712 ± 0.009	0.656 ± 0.012	0.730 ± 0.004	0.921 ± 0.007
12	SimulFG (600)	0.705 ± 0.003	0.658 ± 0.008	0.687 ± 0.006	0.934 ± 0.004
13	SimulFG (650)	0.684 ± 0.004	0.655 ± 0.008	0.700 ± 0.005	0.924 ± 0.004
14	SimulFG (700)	0.693 ± 0.004	0.680 ± 0.009	0.703 ± 0.008	0.941 ± 0.003
15	SimulFG (750)	0.709 ± 0.008	0.649 ± 0.011	0.694 ± 0.005	0.941 ± 0.004
16	SimulFG (800)	0.729 ± 0.005	0.674 ± 0.008	0.723 ± 0.008	0.946 ± 0.005
17	SimulFG (850)	0.721 ± 0.004	0.660 ± 0.010	0.710 ± 0.006	0.951 ± 0.004
18	SimulFG (900)	0.750 ± 0.006	0.703 ± 0.006	0.743 ± 0.004	0.952 ± 0.004
19	SimulFG (950)	0.713 ± 0.005	0.664 ± 0.007	0.706 ± 0.006	0.936 ± 0.004
20	SimulFG (1000)	0.722 ± 0.006	0.670 ± 0.008	0.711 ± 0.003	0.950 ± 0.003
Average		0.704 ± 0.007	0.661 ± 0.010	0.699 ± 0.007	0.920 ± 0.006

Table 5
The detailed experimental results of ranking performance and standard deviation on 20 simultaneous fault datasets.

	Datasets	NNB	FNB	FNB_{ROT}	NNBC
1	SimulFG (50)	0.829 ± 0.010	0.717 ± 0.014	0.809 ± 0.015	0.987 ± 0.005
2	SimulFG (100)	0.879 ± 0.004	0.812 ± 0.008	0.859 ± 0.008	0.982 ± 0.003
3	SimulFG (150)	0.878 ± 0.005	0.799 ± 0.008	0.853 ± 0.004	0.981 ± 0.004
4	SimulFG (200)	0.852 ± 0.011	0.740 ± 0.011	0.819 ± 0.007	0.979 ± 0.004
5	SimulFG (250)	0.850 ± 0.006	0.802 ± 0.012	0.850 ± 0.006	0.986 ± 0.002
6	SimulFG (300)	0.865 ± 0.005	0.783 ± 0.009	0.836 ± 0.005	0.991 ± 0.002
7	SimulFG (350)	0.818 ± 0.004	0.745 ± 0.008	0.799 ± 0.007	0.984 ± 0.003
8	SimulFG (400)	0.852 ± 0.005	0.791 ± 0.009	0.831 ± 0.007	0.986 ± 0.002
9	SimulFG (450)	0.862 ± 0.004	0.818 ± 0.005	0.843 ± 0.004	0.989 ± 0.001
10	SimulFG (500)	0.862 ± 0.005	0.821 ± 0.006	0.850 ± 0.003	0.993 ± 0.001
11	SimulFG (550)	0.862 ± 0.003	0.810 ± 0.006	0.848 ± 0.003	0.986 ± 0.001
12	SimulFG (600)	0.862 ± 0.003	0.783 ± 0.011	0.837 ± 0.004	0.991 ± 0.002
13	SimulFG (650)	0.876 ± 0.003	0.824 ± 0.008	0.870 ± 0.004	0.992 ± 0.001
14	SimulFG (700)	0.838 ± 0.002	0.819 ± 0.005	0.832 ± 0.003	0.991 ± 0.002
15	SimulFG (750)	0.878 ± 0.004	0.840 ± 0.009	0.870 ± 0.005	0.992 ± 0.001
16	SimulFG (800)	0.853 ± 0.002	0.792 ± 0.005	0.836 ± 0.003	0.994 ± 0.001
17	SimulFG (850)	0.845 ± 0.002	0.824 ± 0.004	0.843 ± 0.004	0.995 ± 0.001
18	SimulFG (900)	0.860 ± 0.003	0.833 ± 0.004	0.849 ± 0.003	0.992 ± 0.001
19	SimulFG (950)	0.853 ± 0.002	0.805 ± 0.004	0.832 ± 0.002	0.994 ± 0.001
20	SimulFG (1000)	0.852 ± 0.003	0.806 ± 0.006	0.831 ± 0.003	0.995 ± 0.001
Average		0.856 ± 0.004	0.798 ± 0.008	0.840 ± 0.005	0.989 ± 0.002

with RBFNetwork [26], SMO [26], and J48 [26]. Similarly, we conduct 10-fold cross-validation for 10 times and get the averaged values. The experimental results in terms of classification accuracy, AUC, and PMSE are listed in Fig. 2. From Fig. 2a, we can find that the classification accuracy of our proposed NNBC is superior not only to NNB, FNB, and FNB_{ROT} but also to RBFNetwork, SMO, and J48. In addition, NNBC also obtains the best AUC (Fig. 2b) and PMSE (Fig. 2c) in comparison with NNB, FNB, and FNB_{ROT}.

Then, we compare NNBC with NNB, FNB, and FNB_{ROT} in simultaneous fault diagnosis. The experimental results based on 10 times 10-fold cross-validation in terms of classification accuracy, AUC, and PMSE are summarized in Tables 4–6 respectively. In these tables, SimulFG (n) is the dataset in which all instances are generated with two single faults, where $\{A_1, A_2, A_3, cos(A_1), cos(A_2), sin(A_3)\}$ is the instance belonging to the first class, $\{A_1, A_2, A_3, cos(A_1), sin(A_2), cos(A_3)\}$ is the instance belonging to the second class, and $\{A_1, A_2, A_3, sin(A_1), cos(A_2), cos(A_3)\}$ is the instance belonging to the third class. Each class contains n instances and the size of dataset SinglFG (n) is also $3n$. From Tables 4–6, we can find that the average accuracy, AUC, and PMSE of NNBC are 0.920, 0.989, and 0.136 respectively, which are far superior to those of the other methods. This indicates that NNBC can achieve the most accurate and effective diagnosis for simultaneous faults.

The sensitivity of bandwidth selection on the performances of single and simultaneous fault diagnosis is also empirically validated on two representative datasets SinglFG (100) and SimulFG (100). For each of these two datasets, NNBC is used with the bandwidth parameter h ranging from 0.01 to 5 in step of 0.01. For each h, we evaluate the performances of fault diagnosis in the same way as stated above. The experimental results on SinglFG (100) are summarized in Fig. 3 (Accuracy), Fig. 4 (AUC),

Table 6
The detailed experimental results of estimation quality and standard deviation on 20 simultaneous fault datasets.

	Datasets	NNB	FNB	FNB_{ROT}	NNBC
1	SimulFG (50)	0.435 ± 0.013	0.475 ± 0.018	0.441 ± 0.012	0.196 ± 0.008
2	SimulFG (100)	0.472 ± 0.006	0.637 ± 0.017	0.500 ± 0.008	0.207 ± 0.005
3	SimulFG (150)	0.512 ± 0.012	0.551 ± 0.014	0.508 ± 0.011	0.214 ± 0.008
4	SimulFG (200)	0.439 ± 0.007	0.530 ± 0.010	0.445 ± 0.006	0.163 ± 0.008
5	SimulFG (250)	0.434 ± 0.003	0.481 ± 0.009	0.429 ± 0.005	0.159 ± 0.012
6	SimulFG (300)	0.414 ± 0.003	0.514 ± 0.011	0.432 ± 0.005	0.132 ± 0.004
7	SimulFG (350)	0.408 ± 0.003	0.491 ± 0.006	0.426 ± 0.004	0.128 ± 0.004
8	SimulFG (400)	0.434 ± 0.004	0.510 ± 0.007	0.449 ± 0.003	0.163 ± 0.007
9	SimulFG (450)	0.464 ± 0.006	0.535 ± 0.008	0.477 ± 0.005	0.151 ± 0.005
10	SimulFG (500)	0.399 ± 0.003	0.470 ± 0.008	0.420 ± 0.004	0.141 ± 0.004
11	SimulFG (550)	0.367 ± 0.005	0.453 ± 0.009	0.388 ± 0.006	0.088 ± 0.004
12	SimulFG (600)	0.445 ± 0.003	0.529 ± 0.006	0.456 ± 0.003	0.121 ± 0.003
13	SimulFG (650)	0.391 ± 0.003	0.450 ± 0.008	0.392 ± 0.003	0.090 ± 0.002
14	SimulFG (700)	0.413 ± 0.002	0.482 ± 0.004	0.418 ± 0.003	0.101 ± 0.003
15	SimulFG (750)	0.457 ± 0.004	0.518 ± 0.007	0.472 ± 0.005	0.141 ± 0.003
16	SimulFG (800)	0.406 ± 0.003	0.432 ± 0.006	0.404 ± 0.003	0.099 ± 0.003
17	SimulFG (850)	0.420 ± 0.002	0.507 ± 0.004	0.441 ± 0.002	0.118 ± 0.002
18	SimulFG (900)	0.397 ± 0.003	0.477 ± 0.004	0.420 ± 0.003	0.091 ± 0.001
19	SimulFG (950)	0.395 ± 0.003	0.473 ± 0.005	0.407 ± 0.003	0.108 ± 0.002
20	SimulFG (1000)	0.408 ± 0.002	0.445 ± 0.005	0.400 ± 0.003	0.101 ± 0.002
Average		0.425 ± 0.005	0.498 ± 0.008	0.436 ± 0.005	0.136 ± 0.005

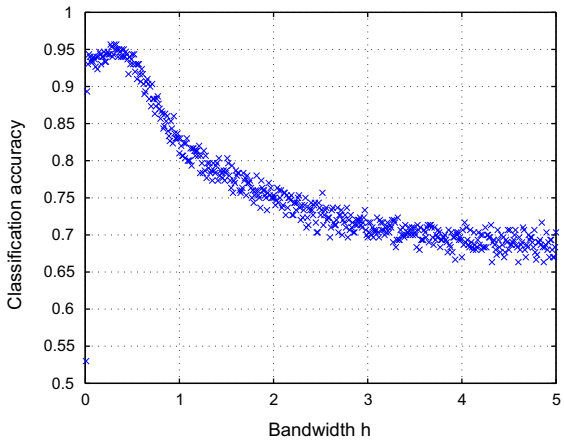


Fig. 3. The sensitivity of bandwidth on classification accuracy in single fault dataset SinglFG (100).

and Fig. 5 (PMSE), and the results on SimulFG (100) are summarized in Fig. 6 (Accuracy), Fig. 7 (AUC), and Fig. 8 (PMSE) respectively. From these figures, we can see that the fault diagnosis based on NNBC is sensitive to the bandwidth selection.

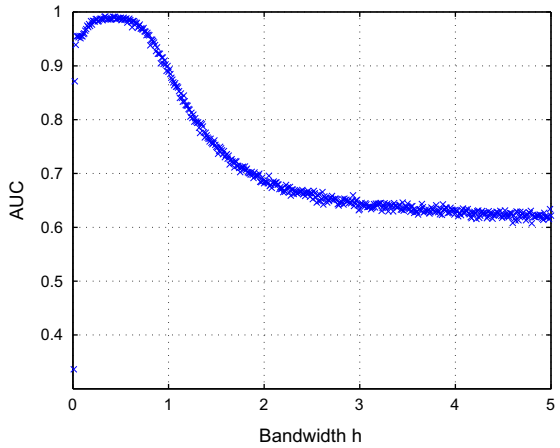


Fig. 4. The sensitivity of bandwidth on AUC in single fault dataset SinglFG (100).

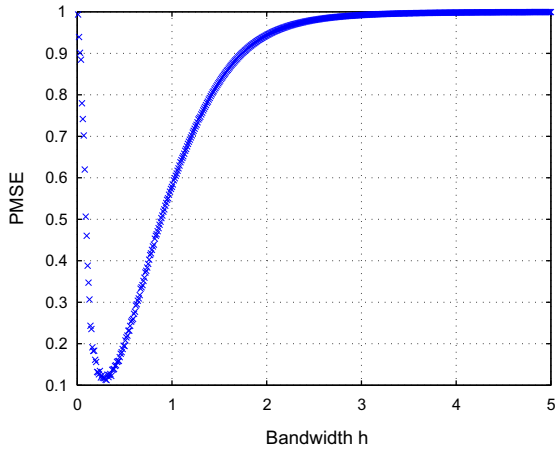


Fig. 5. The sensitivity of bandwidth on PMSE in single fault dataset SinglFG (100).

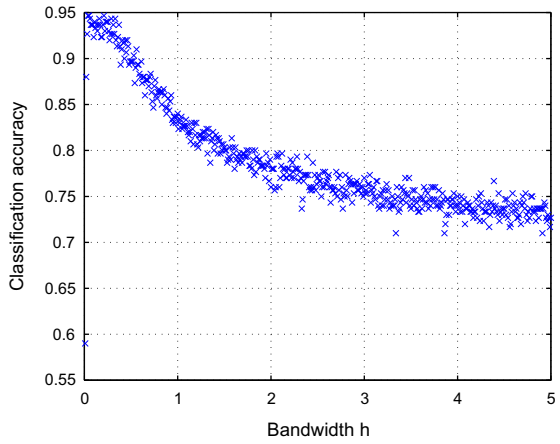


Fig. 6. The sensitivity of bandwidth on classification accuracy in simultaneous fault dataset SimulFG (100).

The optimal bandwidths will lead to the highest accuracies, the highest ranking performances, and the lowest estimation errors. These observations indicate the necessity of the optimal bandwidth selection.

A practical application, known as the chemical process by Watanabe et al. [24] and Eslamloueyan et al. [4], is used for testing our proposed Bayesian model. The schematic diagram of this process is shown in Fig. 9. In this process, the heptane

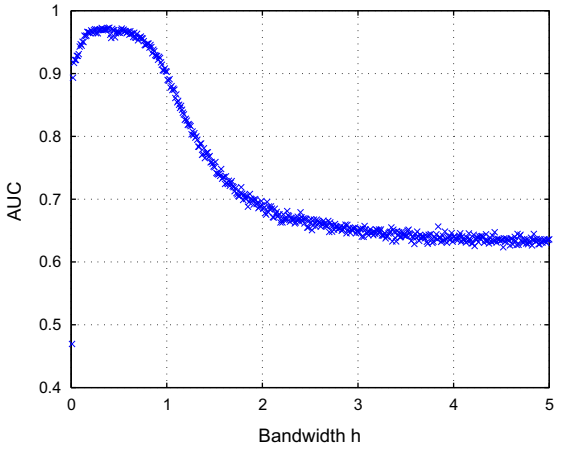


Fig. 7. The sensitivity of bandwidth on AUC in simultaneous fault dataset SimulFG (100).

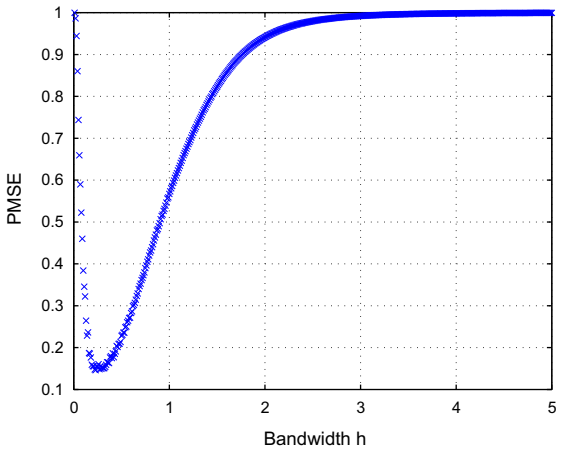


Fig. 8. The sensitivity of bandwidth on PMSE in simultaneous fault dataset SimulFG (100).

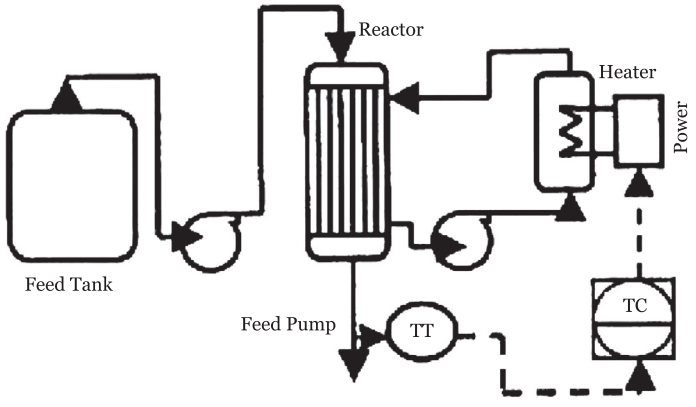


Fig. 9. The schematic diagram of chemical process [24,4].

is converted to toluene in the Reactor which is heated by an electrical Heater installed in the inlet line of Reactor. The detailed descriptions about this chemical process can be found from the Refs. [24,4]. There are seven possible single faults that will occur during the plant operation. These seven single faults are summarized in Table 7, where T_h is the outlet temperature of electrical Heater, T product temperature at the Reactor outlet, S_i integrator output in the PI controller, $C_{C_7H_{16}}$ outlet concentration of heptane, $C_{C_7H_8}$ outlet concentration of toluene, $C^i_{C_7H_{16}}$ heptane concentration in the feed, and T_i Reactor feed temperature.

In Refs. [24,4], the simultaneous fault patterns can be generated with the superposition principle based on the seven single fault patterns. For example, if single faults F_i and F_j are resulted from the deviation vectors of measured variables Δx_i and Δx_j, then, the simultaneous fault $F_i F_j$ is given by $\Delta x_i x_j = \Delta x_i + \Delta x_j$. As used by [24,4], there are 21 different simultaneous faults in every compared dataset, i.e., F_iF_j ($i = 1, 2, \ldots, 6$; $i < j \leqslant 7$). We compare our proposed Bayesian model with NNB, FNB, and FNB_{ROT} on 10 different datasets. Each dataset includes 21 simultaneous faults $F_iF_j(n)$ where $n = 10, 20, \ldots, 100$ denoting the number of patterns belonging to F_iF_j. As proposed by Eslamloueyan et al. [4], we use "deterioration degree" to generate the simultaneous fault patterns based on the pre-defined single faults in Table 7, i.e., $F_i(\alpha)F_j(\beta)$ where α and β are the deterioration degrees which are real numbers in interval $(0, 0.01]$. The experimental results are summarized in Tables 8–10.

From Tables 8–10, we observe that our proposed NNBC obtains the best classification accuracy, ranking performance, and estimation quality. By considering the dependence among conditional attributes, NNBC is indeed feasible and effective when

Table 7
Seven single fault patterns from the plant process [24,4].

Single fault	Measured input pattern						
	ΔT_h	ΔT	ΔS_i	$\Delta C_{C_7H_{16}}$	$\Delta C_{C_7H_8}$	$\Delta C^i_{C_7H_{16}}$	ΔT_i
F_1	−1.85	−0.74	−1.19	5.55	−5.55	0	0
F_2	−0.83	0	−0.83	5.04	−5.04	0	0
F_3	0.90	0	0.90	−5.55	5.55	0	0
F_4	0	0	4.55	0	0	0	0
F_5	3.04	0	3.04	0	0	0	0
F_6	−0.86	0	−0.86	−4.76	−5.23	−10	0
F_7	0.43	0	0.43	0	0	0	−3

Table 8
The detailed experimental results of classification accuracy and standard deviation on 10 simultaneous fault datasets.

	Datasets	NNB	FNB	FNB_{ROT}	NNBC
1	F_iF_j (10)	0.658 ± 0.005	0.718 ± 0.000	0.723 ± 0.004	0.735 ± 0.005
2	F_iF_j (20)	0.753 ± 0.014	0.841 ± 0.004	0.817 ± 0.003	0.882 ± 0.002
3	F_iF_j (30)	0.814 ± 0.004	0.837 ± 0.004	0.830 ± 0.002	0.869 ± 0.006
4	F_iF_j (40)	0.863 ± 0.002	0.897 ± 0.006	0.862 ± 0.005	0.905 ± 0.005
5	F_iF_j (50)	0.734 ± 0.010	0.862 ± 0.002	0.859 ± 0.003	0.892 ± 0.005
6	F_iF_j (60)	0.783 ± 0.003	0.863 ± 0.004	0.869 ± 0.002	0.892 ± 0.004
7	F_iF_j (70)	0.802 ± 0.003	0.889 ± 0.003	0.893 ± 0.003	0.901 ± 0.003
8	F_iF_j (80)	0.846 ± 0.004	0.903 ± 0.004	0.904 ± 0.005	0.900 ± 0.003
9	F_iF_j (90)	0.811 ± 0.004	0.885 ± 0.002	0.884 ± 0.003	0.895 ± 0.002
10	F_iF_j (100)	0.764 ± 0.004	0.882 ± 0.001	0.878 ± 0.005	0.897 ± 0.002
Average		0.783 ± 0.005	0.858 ± 0.003	0.852 ± 0.004	0.877 ± 0.004

Table 9
The detailed experimental results of ranking performance and standard deviation on 10 simultaneous fault datasets.

	Datasets	NNB	FNB	FNB_{ROT}	NNBC
1	F_iF_j (10)	0.983 ± 0.002	0.988 ± 0.001	0.988 ± 0.001	0.992 ± 0.001
2	F_iF_j (20)	0.987 ± 0.001	0.990 ± 0.001	0.988 ± 0.002	0.995 ± 0.001
3	F_iF_j (30)	0.990 ± 0.001	0.996 ± 0.001	0.995 ± 0.001	0.994 ± 0.001
4	F_iF_j (40)	0.985 ± 0.002	0.985 ± 0.001	0.985 ± 0.001	0.991 ± 0.001
5	F_iF_j (50)	0.981 ± 0.001	0.984 ± 0.002	0.985 ± 0.002	0.993 ± 0.001
6	F_iF_j (60)	0.993 ± 0.001	0.994 ± 0.001	0.993 ± 0.001	0.998 ± 0.001
7	F_iF_j (70)	0.988 ± 0.001	0.992 ± 0.001	0.991 ± 0.001	0.996 ± 0.001
8	F_iF_j (80)	0.992 ± 0.000	0.995 ± 0.001	0.995 ± 0.001	0.998 ± 0.000
9	F_iF_j (90)	0.989 ± 0.000	0.996 ± 0.000	0.995 ± 0.001	0.997 ± 0.000
10	F_iF_j (100)	0.993 ± 0.000	0.995 ± 0.000	0.996 ± 0.000	0.998 ± 0.000
Average		0.988 ± 0.001	0.992 ± 0.001	0.991 ± 0.001	0.995 ± 0.001

Table 10
The detailed experimental results of estimation quality and standard deviation on 10 simultaneous fault datasets.

	Datasets	NNB	FNB	FNB_{ROT}	NNBC
1	F_iF_j (10)	0.634 ± 0.012	0.388 ± 0.009	0.475 ± 0.009	0.417 ± 0.004
2	F_iF_j (20)	0.405 ± 0.012	0.291 ± 0.006	0.306 ± 0.003	0.278 ± 0.008
3	F_iF_j (30)	0.456 ± 0.008	0.293 ± 0.006	0.236 ± 0.011	0.261 ± 0.002
4	F_iF_j (40)	0.407 ± 0.005	0.243 ± 0.010	0.236 ± 0.010	0.204 ± 0.006
5	F_iF_j (50)	0.337 ± 0.004	0.237 ± 0.005	0.219 ± 0.005	0.199 ± 0.005
6	F_iF_j (60)	0.340 ± 0.002	0.216 ± 0.003	0.195 ± 0.004	0.182 ± 0.004
7	F_iF_j (70)	0.298 ± 0.006	0.187 ± 0.009	0.180 ± 0.004	0.140 ± 0.005
8	F_iF_j (80)	0.361 ± 0.002	0.209 ± 0.005	0.175 ± 0.005	0.147 ± 0.004
9	F_iF_j (90)	0.290 ± 0.004	0.227 ± 0.005	0.185 ± 0.006	0.171 ± 0.003
10	F_iF_j (100)	0.308 ± 0.004	0.180 ± 0.003	0.158 ± 0.006	0.142 ± 0.001
Average		0.384 ± 0.006	0.247 ± 0.006	0.237 ± 0.006	0.214 ± 0.004

diagnosing the simultaneous faults in the chemical process. In fact, such dependence does occur in the process of plant operation. For example, it is very obvious that there is a dependence between the variables $\Delta C_{C_7H_{16}}$ and $\Delta C_{C_7H_8}$ due to the following reaction [4] which takes place in the Reactor:

$$C_7H_{16(g)} \rightarrow C_7H_{8(g)} + 4H_{2(g)}.$$

Finally, we discuss how to use the single fault instance to diagnose the simultaneous faults. That is to say, when no simultaneous fault instances are available for training and only the single fault instances can be used, how can we use NNB, FNB, and FNB_{ROT} or NNBC to establish the diagnosis model for simultaneous faults. In order to solve this problem, the following framework is constructed (this framework does not require any training instances of simultaneous fault):

(1) Train the base classifier C_{ij} with the single fault instances belonging to the ith and jth ($i, j = 1, 2, 3, i \neq j$) classes, and p_{ij} is the pairwise probability of the ith single fault against the jth single fault for an unknown instance $\tilde{x}$. The probability of instance $\tilde{x}$ containing the ith single fault can be calculated as follows:

$$p_i = \frac{\sum_{i=1}^{3}\sum_{j\neq i}^{3} n_{ij}p_{ij}}{\sum_{i=1}^{3}\sum_{j\neq i}^{3} n_{ij}},$$

where n_{ij} is the number of training instances within the ith and jth labels.

(2) The possible single faults contained in the new instance $\tilde{x}$ can be determined according to the following decision rule:

$$(i,j) = \underset{\substack{i,j=1,2,3 \\ i\neq j}}{\arg\max} \{p_i + p_j\}.$$

Table 11
The results of predicting simultaneous faults in datasets SimulFG (500) based on the models trained on 20 single fault datasets.

	Datasets	NNB	FNB	FNB_{ROT}	NNBC
1	SinglFG (50)	0.692 ± 0.042	0.578 ± 0.041	0.679 ± 0.028	0.858 ± 0.020
2	SinglFG (100)	0.702 ± 0.026	0.595 ± 0.051	0.680 ± 0.029	0.854 ± 0.034
3	SinglFG (150)	0.708 ± 0.027	0.600 ± 0.040	0.689 ± 0.024	0.870 ± 0.028
4	SinglFG (200)	0.701 ± 0.045	0.601 ± 0.045	0.681 ± 0.030	0.886 ± 0.030
5	SinglFG (250)	0.698 ± 0.030	0.626 ± 0.027	0.687 ± 0.030	0.881 ± 0.039
6	SinglFG (300)	0.713 ± 0.032	0.616 ± 0.022	0.700 ± 0.031	0.891 ± 0.030
7	SinglFG (350)	0.707 ± 0.015	0.604 ± 0.038	0.699 ± 0.029	0.902 ± 0.024
8	SinglFG (400)	0.705 ± 0.049	0.604 ± 0.032	0.679 ± 0.056	0.910 ± 0.034
9	SinglFG (450)	0.697 ± 0.043	0.606 ± 0.035	0.688 ± 0.050	0.908 ± 0.030
10	SinglFG (500)	0.723 ± 0.033	0.606 ± 0.039	0.698 ± 0.031	0.905 ± 0.022
11	SinglFG (550)	0.699 ± 0.031	0.623 ± 0.019	0.683 ± 0.033	0.884 ± 0.025
12	SinglFG (600)	0.717 ± 0.029	0.629 ± 0.036	0.709 ± 0.032	0.911 ± 0.030
13	SinglFG (650)	0.726 ± 0.025	0.642 ± 0.028	0.710 ± 0.035	0.909 ± 0.041
14	SinglFG (700)	0.704 ± 0.013	0.615 ± 0.029	0.692 ± 0.014	0.912 ± 0.021
15	SinglFG (750)	0.723 ± 0.017	0.636 ± 0.025	0.702 ± 0.020	0.907 ± 0.024
16	SinglFG (800)	0.723 ± 0.022	0.640 ± 0.039	0.706 ± 0.017	0.919 ± 0.019
17	SinglFG (850)	0.712 ± 0.038	0.621 ± 0.034	0.692 ± 0.036	0.909 ± 0.035
18	SinglFG (900)	0.719 ± 0.025	0.639 ± 0.030	0.698 ± 0.040	0.926 ± 0.024
19	SinglFG (950)	0.727 ± 0.024	0.640 ± 0.029	0.706 ± 0.030	0.935 ± 0.021
20	SinglFG (1000)	0.728 ± 0.028	0.656 ± 0.031	0.711 ± 0.027	0.932 ± 0.025
Average		0.711 ± 0.030	0.619 ± 0.034	0.694 ± 0.031	0.900 ± 0.028

For example, if $p_1 + p_3 > p_1 + p_2$and $p_1 + p_3 > p_2 + p_3$, then the first and third single faults occur simultaneously for the new instance $\tilde{x}$.

By using the above-mentioned strategy, we compare the performances of the four methods in simultaneous fault diagnosis with the models trained on 20 different single fault datasets SinglFG (n), $n = 50, 100, \ldots, 1000$ respectively. For each n, a simultaneous fault dataset SimulFG (500) including 500 simultaneous fault instances is used as the testing set. The experimental results are listed in Table 11. From this table, we can find that the average classification accuracies of NNB, FNB, FNB_{ROT}, and NNBC are 0.711, 0.619, 0.694, and 0.900 respectively. As a result, NNBC still obtains the best diagnosis performance.

In summary, in comparison with NNB, FNB, and FNB_{ROT}, NNBC has the best diagnosis performance in terms of the 3 indices: classification accuracy, ranking and estimation quality. The reason is that, in the NNBC model, the assumption of independence among features is relaxed, and a more accurate estimation of the joint class-conditional p.d.f. can be acquired.

5. Conclusion and future works

In this paper, we propose a new simultaneous fault diagnosis model based on non-naive Bayesian classifier (NNBC). It removes the independence assumption and achieves a more accurate estimation on class-conditional p.d.f.. The comparative results demonstrate that NNBC can obtain the remarkable improvements in the classification accuracy, ranking performance and class-conditional probability estimation. Our scheduled further development in this research topic contains: (i) seeking other application domains for the designed model, such as industrial electronics and control engineering domains; (ii) studying the impact of different bandwidth selection methods on the performance of the model; and (iii) providing the theoretical analysis regarding the sensitivity of optimal bandwidth and finding the locally and globally optimal conditions based on the empirical risk minimization for our proposed NNBC.

Acknowledgements

The authors thank the editors and anonymous reviewers. Their valuable and constructive comments and suggestions helped them in significantly improving this paper. The authors also thank Prof. James Liu for his instructions on the improvement of language quality. This research is supported by the National Natural Science Foundation of China (71371063, 61170040 and 60903089), by the Natural Science Foundation of Hebei Province (F2013201110, F2012201023 and F2011201063), by the Key Scientific Research Foundation of Education Department of Hebei Province (ZD2010139), and by the CRG grant G-YL14 of The Hong Kong Polytechnic University.

References

[1] H.A. Darwish, A.M.I. Taalab, T.A. Kawady, Development and implementation of an ANN-based fault diagnosis scheme for generator winding protection, IEEE Transactions on Power Delivery 16 (2) (2001) 208–214.
[2] S. Deng, S.Y. Lin, W.L. Chang, Application of multiclass support vector machines for fault diagnosis of field air defense gun, Expert Systems with Applications 38 (5) (2011) 6007–6013.
[3] P. Domingos, M. Pazzani, On the optimality of the simple Bayesian classifier under zero-one loss, Machine Learning 29 (2–3) (1997) 103–130.
[4] R. Eslamloueyan, M. Shahrokhi, R. Bozorgmehri, Multiple simultaneous fault diagnosis via hierarchical and single artificial neural networks, Scientia Iranica 10 (3) (2003) 300–310.
[5] D.J. Hand, R.J. Till, A simple generalisation of the area under the ROC curve for multiple class classification problems, Machine Learning 45 (2) (2001) 171–186.
[6] L.X. Jiang, Z.H. Cai, D.H. Wang, H. Zhang, Bayesian citation-KNN with distance weighting, International Journal of Machine Learning and Cybernetics (in press). doi:http://dx.doi.org/10.1007/s13042-013-0152-x.
[7] G.H. John, P. Langley, Estimating continuous distributions in Bayesian classifiers, in: Proceedings of UAI'95, Quebec, 1995, pp. 338–345.
[8] M. Kobos, Combination of independent kernel density estimators in classification, in: Proceedings of IMCSIT'09, Mra˛gowo, 2009, pp. 57–63.
[9] A. Lemos, W. Caminhas, F. Gomide, Adaptive fault detection and diagnosis using an evolving fuzzy classifier, Information Sciences 220 (2013) 64–85.
[10] Z.L. Li, X.B. Xu, C.L. Wen, A new approach of simultaneous faults diagnosis based on random sets and DSmT, Journal of Electronics (China) 26 (1) (2009) 24–30.
[11] B. Liu, Y. Yang, G.I. Webb, J. Boughton, A comparative study of bandwidth choice in kernel density estimation for naive bayesian classification, Lecture Notes in Computer Science 5476 (2009) 302–313.
[12] M.S. Mahmoud, H.M. Khalid, Expectation maximization approach to data-based fault diagnostics, Information Sciences 235 (2013) 80–96.
[13] J. Martinka, K. Kozlíková, Dependence of the vulnerability index on the heart cycle length, Physiological Research 56 (S1) (2007) 129–132.
[14] T. Mitchell, Machine Learning, McGraw Hill, 1997.
[15] V. Muralidharan, V. Sugumaran, A comparative study of Naive Bayes classifier and Bayes net classifier for fault diagnosis of monoblock centrifugal pump using wavelet analysis, Applied Soft Computing 12 (8) (2012) 2023–2029.
[16] S. Nandi, H.A. Toliyat, X.D. Li, Condition monitoring and fault diagnosis of electrical motors-a review, IEEE Transactions on Energy Conversion 20 (4) (2005) 719–729.
[17] H. Peng, J. Wang, M.J. Perez-Jimenez, H. Wang, J. Shao, T. Wang, Fuzzy reasoning spiking neural P system for fault diagnosis, Information Sciences 235 (2013) 106–116.
[18] Y. Qian, X.X. Li, Y.R. Jiang, Y.Q. Wen, An expert system for real-time fault diagnosis of complex chemical processes, Expert Systems with Applications 24 (4) (2003) 425–432.
[19] D.W. Scott, Multivariate Density Estimation: Theory, Practice, and Visualization, John Wiley and Sons, Inc., 1992.
[20] N.C. Tsai, Y.H. King, R.M. Lee, Fault diagnosis for magnetic bearing systems, Mechanical Systems and Signal Processing 23 (4) (2009) 1339–1351.
[21] UCI Machine Learning Repository. <http://archive.ics.uci.edu/ml/>.

[22] R. Vullings, B. De Vries, J.W.M. Bergmans, An adaptive Kalman filter for ECG signal enhancement, IEEE Transactions on Biomedical Engineering 58 (4) (2011) 1094–1103.
[23] M.P. Wand, M.C. Jones, Kernel Smoothing, Chapman and Hall, 1995.
[24] K. Watanabe, S. Hirota, L. Hou, D.M. Himmeblau, Diagnosis of multiple simultaneous fault via hierarchical artificial neural networks, American Institute of Chemical Engineers Journal 40 (5) (1994) 839–848.
[25] G.I. Webb, J.R. Boughton, F. Zheng, K.M. Ting, H. Salem, Learning by extrapolation from marginal to full-multivariate probability distributions: decreasingly naive Bayesian classification, Machine Learning 86 (2) (2012) 233–272.
[26] I.H. Witten, E. Frank, Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufmann, 2005.
[27] J.Z. Xiao, H.R. Wang, X.C. Yang, Z. Gao, Multiple faults diagnosis in motion system based on SVM, International Journal of Machine Learning and Cybernetics 3 (1) (2012) 77–82.
[28] Z. Yang, W.H. Tang, A. Shintemirov, Q.H. Wu, Association rule mining-based dissolved gas analysis for fault diagnosis of power transformers, IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews 39 (6) (2009) 597–610.
[29] I. Yélamos, M. Graells, L. Puigjaner, G. Escudero, Simultaneous fault diagnosis in chemical plants using a multilabel approach, American Institute of Chemical Engineers Journal 53 (11) (2007) 2871–2884.
[30] F. Zheng, G.I. Webb, P. Suraweera, L.G. Zhu, Subsumption resolution: an efficient and effective technique for semi-naive Bayesian learning, Machine Learning 87 (1) (2012) 93–125.