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

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).

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:

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

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):

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