2014 International Joint Conference on Neural Networks (IJCNN)
July 6-11, 2014, Beijing, China

Hybrid Classification with Partial Models

Bo Tang, Quan Ding, Haibo He, and Steven Kay

***Abstract*— The parametric classifiers trained with the Bayesian rule are usually more accurate than the non-parametric classifiers such as nearest neighbors, neural network and support vector machine, when the class-conditional densities of distribution models are known except for some of their parameters and the training data is abundant. However, the parametric classifiers would perform poorly if these class-conditional densities are unknown and the assumed distribution models are inaccurate. In this paper, we propose a hybrid classification method for the data with partially known distribution models where only the distribution models of some classes are known. For this partial models case, the proposed hybrid classifier makes the best use of knowledge of known distribution models with Bayesian interference, while both purely parametric and non-parametric classifiers would lose a specific predictive capacity for classification. Theoretical proofs and experimental results show that the proposed hybrid classifier has much better performance than these purely parametric and non-parametric classifiers for the data with partial models.**

I. Introduction

In classification problems, generative and discriminative classification methods are two well-known classifiers. The generative classification approach learns the joint probability $p(\mathbf{x}, y)$, where $\mathbf{x}$ is the input data vector and y is the corresponding class label, and makes a classification decision based on the posterior probability $p(y|\mathbf{x})$ which is calculated with Bayesian rule in Eq. (1). In contrast, the alternative discriminative approach directly model the posterior probability $p(y|\mathbf{x})$ from the input training data set. Both generative and discriminative methods make predictive decision based on the posterior probability and choose the most likely class label for the input data $\mathbf{x}$. In practice, both of them provide outstanding performance and are widely used for various classification problems.

$$p(y|\mathbf{x}) = \frac{p(y)p(\mathbf{x}|y)}{\sum_{i=1}^{N} p(C_i)p(\mathbf{x}|C_i)} \tag{1}$$

To make a predictive decision, the generative approach needs to model the class-conditional density $p(\mathbf{x}|y)$ which can be obtained for prior knowledge or learnt from the data, while the discriminative approach need to model the posterior probability $p(y|\mathbf{x})$ which has to be learnt from data directly. In generative approach, the class-conditional density function $p(\mathbf{x}|y)$ usually has a parametric form, for example, one common assumption of $p(\mathbf{x}|y)$ is Gaussian or Gaussian mixture model for which their parameters, such as mean vector and covariance matrix, are unknown and need to be estimated. In discriminative approach, the posterior probability $p(y|\mathbf{x})$ is usually modeled with a non-parametric form estimated from the data directly, such as nearest neighbors, neural network, and support vector machine, to name a few.

Given a specific classification problem, if the class-conditional density functions $p(\mathbf{x}|y)$ of all classes are known, a parametric classifier with Bayes' theorem achieves the best performance, while non-parametric classifiers would lose some performance because they do not make any use of these knowledge. Even when $p(\mathbf{x}|y)$ is known to be within a family of probability density functions (PDFs) parameterized by some unknown parameters, it also has been proved that the parametric classifiers can asymptotically approximate the optimal classifier replacing the unknown parameters with their maximum likelihood estimations (MLEs) [1]. In many practical cases, the forms of class-conditional density $p(\mathbf{x}|y)$ are difficult to obtain, and thus a parametric classifier usually lays down a strong assumption of the underlying data model. It would perform poorly if such an assumption is inappropriate. For a non-parametric classifier, there is no such assumption or no density estimation for class-conditional distributions. In this paper, to the best of our knowledge, we are the first time to address the classification problem when the distribution models of some classes are available, i.e. the form of $p(\mathbf{x}|y)$ is obtainable except for their parameters, while the distribution models of other classes are completely unknown. For the partially known distribution models, both purely parametric and non-parametric classifiers would lose a specific predictive capacity for classification. To solve this problem, we propose a hybrid classification method in which we combine both parametric and non-parametric classifiers to build a powerful decision maker.

The remaining paper is organized as follows. We first present the related works in Section II. In Section III, we formulate the problem of partially known distribution models, propose a hybrid classification method, and theoretically prove its effectiveness. In Section IV, we apply the proposed classification method on synthetic Gaussian distribution data and power quality disturbance data and compare the performance with purely non-parametric classifiers. Finally, a conclusion is given in Section V.

II. Related Works

The comparison of the generative and discriminative approaches is a long-standing debate in machine learning area. It is still hard to give a right answer as both ways of predicating class label are based on the posterior probability $p(y|\mathbf{x})$ [2]. In many practical cases, the conditional class

Bo Tang, Haibo He, and Steven Kay are with the Department of Electrical, Computer, and Biomedical Engineering, University of Rhode Island, Kingston, RI, USA (email: {btang, he, and kay}@ele.uri.edu), Quan Ding is with the Department of Physiological Nursing, Univerisity of California, San Francisco, CA, USA (email: Quan.Ding@nursing.ucsf.edu).

This work was supported from National Science Foundation (NSF) under grant CAREER ECCS 1053717, the Army Research Office under grant W911NF-12-1-0378, and the NSF-DFG Collaborative Research on Autonomous Learning, a supplement grant to NSF CNS 1117314

978-1-4799-1484-5/14/$31.00 ©2014 IEEE

density is unknown and need to be estimated from data for a generative approach. To estimate the conditional class density, the assumption of distribution model for each class is made, such as naive Bayes where it assumes that all classes have a Gaussian distribution. Ng and Jordan in [3] compared the predictive performance of generative naive Bayes classifier and discriminative logistic regression algorithm. They showed that the discriminative methods would perform better than the naive Bayes classifier over several real-life datasets in which there are enough training data. They also presented that the naive Bayes classifier would outperform the discriminative approaches if the size of training data is small.

To take advantages of both approaches, several hybrid classification methods have been proposed [4]–[9]. Tong and Koller in [4] proposed a restricted Bayes classifier in which a Bayes optimal classifier is built by minimizing the estimated Bayes error within a certain class. The proposed hybrid classifier would increase classification performance when training data set contains samples with missing feature values. In [5], every input feature vector is divided into several subvectors with which different hypotheses based on naive Bayes classifier are built. A final discriminative hypothesis is built by combining these subgenerative models. In [6], the authors proposed a hybrid approach to semi-supervised classification in which generative and bias correction models are combined with the maximum entropy principle. It is worth noticing that, for these hybrid classification methods, the conditional class density for each class could be unknown and they made a strong assumption of the distributions under each class. Unlike the existing hybrid classification methods, in this paper, we consider the classification problem with partially known distribution models, i.e. only distribution models of some classes are known.

III. Hybrid Classification Methods

For the classification problems in which only the distribution models under some classes are known, both parametric and non-parametric classifiers would lose some predictive capacity. In this partially known distribution model, we propose a hybrid classification method to make the best use of the knowledge of known distributions and build a powerful decision maker. The underlying idea of hybrid classification method is that we separate the classification into two steps: grouping identification and sub-classifications. Given a new test data to be classified, in the grouping identification step, we first classify the test data $\mathbf{x}$ into two groups: the group $C_1^{'}$ of known distribution models and the group $C_2^{'}$ of unknown distribution models. For the data belongs to $C_1^{'}$, we apply a parametric classifier $h_1(\mathbf{x})$ to make a final classification among the classes whose distribution models are known. In contrast, if the data belongs to $C_2^{'}$, we apply a non-parametric classifier $h_2(\mathbf{x})$ to make a final classification among the classes whose distribution models are unknown. For the parametric classifier $h_1(\mathbf{x})$, we calculate the posterior probability $p(y|\mathbf{x})$ in Eq. (1) and replace the parameter $\boldsymbol{\theta}$ in the class-conditional density $p(\mathbf{x}|y,\boldsymbol{\theta})$ by its MLE $\hat{\boldsymbol{\theta}}$. For

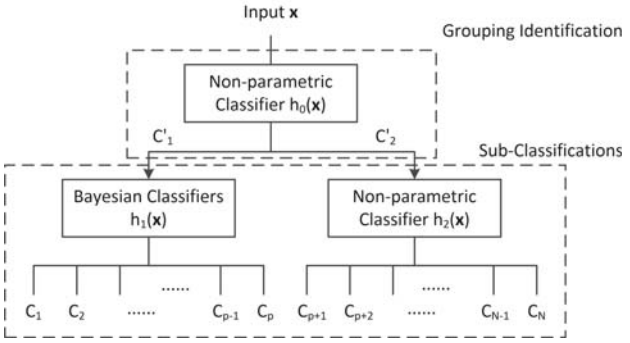

Fig. 1. The framework of hybrid classification method

the non-parametric classifier $h_2(\mathbf{x})$, common non-parametric classifiers, such as nearest neighbor, neural network or support vector machine (SVM), could be used to build a classification hypothesis. The general framework of hybrid classification method is shown in Fig. 1.

Consider the classification problem with N classes, let $\mathcal{Y} = \{C_1, C_2, \cdots, C_N\}$ classes, we assume the first p class models are known, i.e. $p(\mathbf{x}|C_1,\boldsymbol{\theta}_1), p(\mathbf{x}|C_2,\boldsymbol{\theta}_2), \cdots, p(\mathbf{x}|C_p,\boldsymbol{\theta}_p)$ are known except that the parameters $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \cdots, \boldsymbol{\theta}_p$ are unknown, and the remaining $N-p$ class models are unknown. Let there are n training data $\mathcal{X} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \cdots, (\mathbf{x}_n, y_n)\}$ with their corresponding class labels. In the grouping identification step, we firstly identify which type of model it has: known or unknown data model. To do that, we group the training data into two categories: the group $C_1^{'} = C_1 \cup C_2 \cup \cdots \cup C_p$ including these classes with known data model, and the group $C_2^{'} = C_{p+1} \cup C_{p+2} \cup \cdots \cup C_N$ including these classes with unknown data model. To be clearly understood, in hybrid classification method, we refer the classes $C_1, C_2, \cdots, C_p$ in $C_1^{'}$ and the classes $C_{p+1}, C_{p+2}, \cdots, C_N$ in $C_2^{'}$ as classes, and $C_1^{'}$ and $C_2^{'}$ as groups. For the new two groups, we build a parametric or non-parametric classifier $h_0(\mathbf{x})$ in which $p(C_1^{'}|\mathbf{x})$ and $p(C_2^{'}|\mathbf{x})$ could be obtained for any input data vector $\mathbf{x}$. Given the test data $\mathbf{x}$, if $p(C_1^{'}|\mathbf{x}) > p(C_2^{'}|\mathbf{x})$, $\mathbf{x}$ are considered as from the group $C_1^{'}$ with known data model. Otherwise, $\mathbf{x}$ are considered as from the group $C_2^{'}$ with unknown data model. After that, the sub-classification step is employed to further decide which class it belongs to. For the group $C_1^{'}$ with p classes, the parameter for each class could be estimated by the MLEs $\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2, \cdots, \hat{\boldsymbol{\theta}}_p$. Hence, the class-conditional density for each class $p(\mathbf{x}|C_1), p(\mathbf{x}|C_2), \cdots, p(\mathbf{x}|C_N)$ could be replaced as $p(\mathbf{x}|C_1,\hat{\boldsymbol{\theta}}_1), p(\mathbf{x}|C_2,\hat{\boldsymbol{\theta}}_2), \cdots, p(\mathbf{x}|C_N,\hat{\boldsymbol{\theta}}_N)$. If the $\mathbf{x}$ is classified as $C_1^{'}$ in the grouping identification step, we decide the $\mathbf{x}$ belongs to the class i as following

$$\begin{aligned} h_1(\mathbf{x}) &= \underset{i=1,2,\cdots,p}{\arg\max}\, p(C_i|C_1^{'}) p(\mathbf{x}|C_i,\hat{\boldsymbol{\theta}}_i, C_1^{'}) \\ &= \underset{i=1,2,\cdots,p}{\arg\max}\, p(C_i|C_1^{'}) p(\mathbf{x}|C_i,\hat{\boldsymbol{\theta}}_i) \end{aligned} \tag{2}$$

The above equations are equal because $C_i \subset C_1^{'}$. For the

group $C_2^{'}$ with $N-p$ classes, we train another non-parametric classifier $h_2(\mathbf{x})$ to further identify the class it belongs to if $\mathbf{x}$ is classified as $C_2^{'}$ in the grouping identification step. The hybrid classification method is summarized in Table I.

TABLE I
THE HYBRID CLASSIFICATION METHOD

Input:
Training data set $\mathcal{X} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \cdots, (\mathbf{x}_n, y_n)\}$, where $y_i \in \mathcal{Y} = \{C_1, C_2, \cdots, C_N\}$
The first p class's distributions $p(\mathbf{x}|C_1, \boldsymbol{\theta}_1), \cdots, p(\mathbf{x}|C_p, \boldsymbol{\theta}_p)$, where $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \cdots, \boldsymbol{\theta}_p$ are unknown.
Training Step:
1. Estimate the unknown parameters $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \cdots, \boldsymbol{\theta}_p$ as $\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2, \cdots, \hat{\boldsymbol{\theta}}_p$ with MLE.
2. Group the data from p classes as group $C_1^{'}$, and the remaining data as group $C_2^{'}$.
3. Build a non-parametric hypothesis $h_0(\mathbf{x})$ for $C_1^{'}$ and $C_2^{'}$.
4. Build a Bayesian rule $h_1(\mathbf{x})$ for the p classes within $C_1^{'}$.
5. Build a non-parametric hypothesis $h_2(\mathbf{x})$ for the $N-p$ classes within $C_2^{'}$.
Testing Step: Given a test data $\mathbf{x}_z$,
1. Apply $\mathbf{x}_z$ to $h_0(\mathbf{x})$
2. If $h_0(\mathbf{x}_z) == C_1^{'}$, apply $h_1(\mathbf{x}_z)$ for a final classification: $h_1(\mathbf{x}_z) = 1, 2, \cdots, p$.
3. Else, apply $h_2(\mathbf{x}_z)$ for a final classification: $h_2(\mathbf{x}_z) = p+1, p+2, \cdots, N$.

Definition 1: The Bayes' minimum error for this N classes classification problem can be defined by

$$\begin{aligned} R = p(\text{error}) &= \sum_{i=1}^{N} p(\text{error}|C_i)p(C_i) \\ &= \sum_{i=1}^{N} p(C_i) \int_{\Omega - \Omega_i} p(\mathbf{x}|C_i)d\mathbf{x} \end{aligned} \tag{3}$$

where the integral is taken over $\Omega - \Omega_i$, the region of measurement space outside Ω_i, where $\Omega = \sum_{j=1}^{N} \Omega_j$, and $\Omega - \Omega_i$ means the complement operator, i.e. $\sum_{i=j, j\neq i}^{N} \Omega_j$. The region Ω_i is the classification region for which $p(C_i|\mathbf{x})$ is the largest over all classes.

Similarly, the Bayes' minimum error for the identification classifier $h_0(\mathbf{x})$ could be defined by R_0^* as

$$\begin{aligned} R_0^* &= \sum_{i=1}^{2} p(\text{error}|C_i^{'})p(C_i^{'}) \\ &= p(C_1^{'}) \int_{\Omega_2^{'}} p(\mathbf{x}|C_1^{'})d\mathbf{x} + p(C_2^{'}) \int_{\Omega_1^{'}} p(\mathbf{x}|C_2^{'})d\mathbf{x} \end{aligned} \tag{4}$$

where $\Omega_1^{'} = \sum_{i=1}^{p} \Omega_i$ and $\Omega_2^{'} = \sum_{i=p+1}^{N} \Omega_i$.

Theorem 1: The Bayes' minimum error R_0^* of grouping identification classifier for N_g groups classification problem is less than the Bayes' minimum error R for N classes classification problem.

PROOF. Considering the case with $N_g = 2$, according to the Bayesian theorem, we have

$$\begin{aligned} R_0^* &= p(C_1^{'}) \int_{\Omega_2^{'}} p(\mathbf{x}|C_1^{'})d\mathbf{x} + p(C_2^{'}) \int_{\Omega_1^{'}} p(\mathbf{x}|C_2^{'})d\mathbf{x} \\ &= \int_{\Omega_2^{'}} p(C_1^{'}|\mathbf{x})p(\mathbf{x})d\mathbf{x} + \int_{\Omega_1^{'}} p(C_2^{'}|\mathbf{x})p(\mathbf{x})d\mathbf{x} \\ &= \int_{\Omega_2^{'}} p(C_1 \cup C_2 \cup \cdots \cup C_p|\mathbf{x})p(\mathbf{x})d\mathbf{x} + \\ &\quad \int_{\Omega_1^{'}} p(C_{p+1} \cup C_{p+2} \cup \cdots \cup C_N|\mathbf{x})p(\mathbf{x})d\mathbf{x} \end{aligned} \tag{5}$$

Because

$$p(x \cup y) \leq p(x) + p(y) \tag{6}$$

Eq. (5) can be written as

$$\begin{aligned} R_0^* &= \int_{\Omega_2^{'}} p(C_1 \cup C_2 \cup \cdots \cup C_p|\mathbf{x})p(\mathbf{x})d\mathbf{x} + \\ &\quad \int_{\Omega_1^{'}} p(C_{p+1} \cup C_{p+2} \cup \cdots \cup C_N|\mathbf{x})p(\mathbf{x})d\mathbf{x} \\ &\leq \sum_{i=1}^{p} p(C_i) \int_{\Omega_2^{'}} p(\mathbf{x}|C_i)d\mathbf{x} + \sum_{i=p+1}^{N} p(C_i) \int_{\Omega_1^{'}} p(\mathbf{x}|C_i)d\mathbf{x} \end{aligned} \tag{7}$$

Since $\Omega_1^{'} = \sum_{i=1}^{p} \Omega_i$ and $\Omega_2^{'} = \sum_{i=p+1}^{N} \Omega_i$, we have

$$\begin{aligned} R_0^* &\leq \sum_{i=1}^{p} p(C_i) \int_{\Omega_i} p(\mathbf{x}|C_i)d\mathbf{x} + \sum_{i=p+1}^{N} p(C_i) \int_{\Omega_i} p(\mathbf{x}|C_i)d\mathbf{x} \\ &= \sum_{i=1}^{N} p(C_i) \int_{\Omega_i} p(\mathbf{x}|C_i)d\mathbf{x} = R \end{aligned} \tag{8}$$

Proof done.

We can also easily extend Theorem 1 to the case that the training data are grouped into more than two classes, i.e. $N_g > 2$. Theorem 1 indicates that we can obtain better classification performance if we group multiple classes into one class. This also can be explained with the straightforward way: the misclassifications between the grouping classes are gone in the new constructed classes.

We further examine the Bayes' minimum error of hybrid classification method in comparison with the original one without grouping identification and sub-classifications.

Definition 2: The Bayes' minimum error of hybrid classification method with any classifier for N classes classification problem can be defined by

$$\begin{aligned} R_1^* = p(\text{error}) &= \sum_{i=1}^{N} p(\text{error}|C_i)p(C_i) \\ &= \sum_{i=1}^{p} p(C_i)p(\underbrace{\Omega_2^{'} \cup (\Omega_1^{'} \cap (\Omega_1^{'} - \Omega_i))}_{\text{error events}}|C_i) + \\ &\quad \sum_{i=p+1}^{N} p(C_i)p(\underbrace{\Omega_1^{'} \cup (\Omega_2^{'} \cap (\Omega_2^{'} - \Omega_i))}_{\text{error events}}|C_i) \end{aligned} \tag{9}$$

where $p(\Omega_j^{'}|C_i)$ denotes the probability that $\mathbf{x}$ which belongs to C_i class locates in the area $\Omega_j^{'}$ with the following form:

$$p(\Omega_j^{'}|C_i) = \int_{\Omega_j^{'}} p(\mathbf{x}|C_i)d\mathbf{x} \tag{10}$$

Theorem 2: The hybrid classification method with any classifier $h_0(\mathbf{x}), h_1(\mathbf{x})$, and $h_2(\mathbf{x})$ has the same Bayes' minimum error as the original one, i.e. $R_1^* = R$.

PROOF. We first consider two groups hybrid classification, and the proof can be easily extended to any N_g groups hybrid classification. Because Bayes' minimum error is the optimal error rate which only depends on the underlying distributions of each class, it provides the upper bound of error rate for any classifier $h_0(\mathbf{x}), h_1(\mathbf{x})$, and $h_2(\mathbf{x})$ in hybrid classification method. In R_1^*, for each sub-class which belongs to $C_1^{'}$, we have the following error probability

$$\begin{aligned} p(\text{error}|C_1) &= p(\Omega_2^{'} \cup (\Omega_1^{'} \cap (\Omega_1^{'} - \Omega_1))|C_1) \\ &= p(\Omega_2^{'}|C_1) + p(\Omega_1^{'} \cap \Omega_1^{'} - \Omega_1|C_1) \\ &= p(\Omega_2^{'}|C_1) + p(\Omega_1^{'} - \Omega_1|C_1) \\ &= p(\Omega - \Omega_1|C_1) = \int_{\Omega-\Omega_1} p(\mathbf{x}|C_1)d\mathbf{x} \end{aligned} \tag{11}$$

In the same way, for each sub-class which belongs to $C_2^{'}$, we have the following error probability

$$p(\text{error}|C_2) = \int_{\Omega-\Omega_2} p(\mathbf{x}|C_2)d\mathbf{x} \tag{12}$$

Hence, according to Eq. (3), we have

$$\begin{aligned} R_1^* &= \sum_{i=1}^{N} p(\text{error}|C_i)p(C_i) \\ &= \sum_{i=1}^{N} p(C_i) \int_{\Omega-\Omega_i} p(\mathbf{x}|C_i)d\mathbf{x} = R \end{aligned} \tag{13}$$

Proof done.

Theorem 2 demonstrates that the proposed hybrid classification method has the same lower bound of error rate which always is considered as the optimal error rate.

Definition 3: For a classification hypothesis h, the generalization error is defined as

$$R_g(h) = \Pr[h(\mathbf{x}) \neq y] = E[1_{h(\mathbf{x}_i)\neq y_i}] \tag{14}$$

where 1_w is the indicator function of the event w, $h(\mathbf{x}_i)$ denotes the label which is assigned to the data $\mathbf{x}_i$ in classifier, and y_i is the true class label it belongs to.

Theorem 3: The hybrid classification method provides lower generalization error rate $R_g^*(h)$ than purely non-parametric classifiers $R_g(h)$ which don't make any use of the knowledge of known distribution models, i.e. $R_g^*(h) \leq R_g(h)$.

PROOF. In hybrid classification method, the classifier includes two steps: identification and sub-classification. Hence, the hypothesis $h(\mathbf{x})$ in Eq. (14) is the combination of identification decision $h_0(\mathbf{x})$ and sub-classification decision $h_1(\mathbf{x})$ or $h_2(\mathbf{x})$. We have,

$$\begin{aligned} R_g^*(h) &= E[1_{h_0(\mathbf{x})\neq C_1^{'}||(h_0(\mathbf{x})=C_1^{'}\wedge h_1(\mathbf{x})\neq y))}|C_1^{'}] \\ &\quad + E[1_{h_0(\mathbf{x})\neq C_2^{'}||(h_0(\mathbf{x})=C_2^{'}\wedge h_2(\mathbf{x}_i)\neq y_i))}|C_2^{'}] \\ &= E(1_{h_0(\mathbf{x})\neq C_1^{'}}|C_1^{'}] + E[1_{(h_0(\mathbf{x})=C_1^{'}\wedge h_1(\mathbf{x})\neq y_i))}|C_1^{'}] \\ &\quad + E[(1_{h_0(\mathbf{x})\neq C_2^{'}}|C_2^{'}] + E[1_{(h_0(\mathbf{x})=C_2^{'}\wedge h_2(\mathbf{x})\neq y_i))}|C_2^{'}] \end{aligned} \tag{15}$$

For purely non-parametric classifiers with identification and sub-classification steps, the available known information of class models are disregarded. However, in hybrid classification method, we build the parametric classifier $h_1(\mathbf{x})$ for sub-classification based on the known class models, which indicates

$$\Pr(h_1(\mathbf{x}) \neq y) \leq \Pr(h_s(\mathbf{x}) \neq y) \tag{16}$$

where $h_s(\mathbf{x})$ is the sub-classification classifier in purely non-parametric classifiers which don't make use of the available known data model. We have

$$\begin{aligned} R_g^*(h) \leq & E[1_{h_0(\mathbf{x})\neq C_1^{'}}|C_1^{'}] + E[1_{(h_0(\mathbf{x})=C_1^{'}\wedge h_s(\mathbf{x})\neq y_i))}|C_1^{'}] \\ & + E[(1_{h_0(\mathbf{x})\neq C_2^{'}}|C_2^{'}] + E[1_{(h_0(\mathbf{x})=C_2^{'}\wedge h_2(\mathbf{x})\neq y_i))}|C_2^{'}] \\ & = R_g(h) \end{aligned} \tag{17}$$

IV. SIMULATIONS

A. Multivariate Gaussian Distributions

First, we consider four classes classification problem ($N = 4$) in which all of them C_1, C_2, C_3 and C_4 satisfy standard bivariate Gaussian distributions. The simulation parameters are shown in Table II, where $\mathbf{I}_{2\times 2}$ is the 2×2 identity matrix. The snapshot of data distribution is shown in Fig. 2. For the partial model, we further assume that the distribution of C_1 and C_2 classes are known except their distribution parameters, while the distribution of C_3 and C_4 are unknown. The parameters of distribution for class C_1 and C_2 are estimated with Expectation-Maximization (EM) method. In spite of the available known information of two class's distributions, it is hard to embedded these information to improve predictive capability for non-parametric classifiers. We compare the performance of proposed hybrid classifiers with purely non-parametric classifiers including nearest neighbors and neural network, and with the optimal maximum a posteriori probability (MAP) rule in which all class's distributions are fully known. Given the distribution of all the classes, the MAP rule is the most optimal classifier which provides the upper bound classification accuracy while the size of training data goes to infinity. In the MAP rule, given the same prior distribution for each class, the testing data $\mathbf{x}_z$ is assigned the class label when the following target function is maximized over i

$$\begin{aligned} \ln p(\mathbf{x}|C_i) = -\frac{1}{2}\sum_{k=1}^{3} \boldsymbol{\alpha}_{i,k}[(\mathbf{x}-\boldsymbol{\mu}_{i,k})^T \boldsymbol{\Sigma}_{i,k}^{-1}(\mathbf{x}-\boldsymbol{\mu}_{i,k}) \\ + \ln\det \boldsymbol{\Sigma}_{i,k}] \end{aligned} \tag{18}$$

TABLE II

THE PARAMETERS OF DISTRIBUTION FOR CLASS C_1, C_2, C_3 AND C_4

Class	Distribution	Parameters
C_1	$\sum_{i=1}^{3}\alpha_{1,i}\mathcal{N}(\boldsymbol{\mu}_{1,i},\boldsymbol{\Sigma}_{1,i})$	$\boldsymbol{\mu}_{1,1}=[4\quad 0]$, $\boldsymbol{\Sigma}_{1,1}=\mathbf{I}_{2\times 2}$ $\boldsymbol{\mu}_{1,2}=[0\quad 4]$, $\boldsymbol{\Sigma}_{1,2}=2\mathbf{I}_{2\times 2}$ $\boldsymbol{\mu}_{1,3}=[0\quad -4]$, $\boldsymbol{\Sigma}_{1,3}=3\mathbf{I}_{2\times 2}$
C_2	$\sum_{i=1}^{3}\alpha_{2,i}\mathcal{N}(\boldsymbol{\mu}_{2,i},\boldsymbol{\Sigma}_{2,i})$	$\boldsymbol{\mu}_{2,1}=[0\quad 4]$, $\boldsymbol{\Sigma}_{2,1}=\mathbf{I}_{2\times 2}$ $\boldsymbol{\mu}_{2,2}=[4\quad 0]$, $\boldsymbol{\Sigma}_{2,2}=2\mathbf{I}_{2\times 2}$ $\boldsymbol{\mu}_{2,3}=[-4\quad 0]$, $\boldsymbol{\Sigma}_{2,3}=3\mathbf{I}_{2\times 2}$
C_3	$\sum_{i=1}^{3}\alpha_{3,i}\mathcal{N}(\boldsymbol{\mu}_{3,i},\boldsymbol{\Sigma}_{3,i})$	$\boldsymbol{\mu}_{3,1}=[-4\quad 0]$, $\boldsymbol{\Sigma}_{3,1}=\mathbf{I}_{2\times 2}$ $\boldsymbol{\mu}_{3,2}=[0\quad -4]$, $\boldsymbol{\Sigma}_{3,2}=2\mathbf{I}_{2\times 2}$ $\boldsymbol{\mu}_{3,3}=[0\quad 4]$, $\boldsymbol{\Sigma}_{3,3}=3\mathbf{I}_{2\times 2}$
C_4	$\sum_{i=1}^{3}\alpha_{4,i}\mathcal{N}(\boldsymbol{\mu}_{4,i},\boldsymbol{\Sigma}_{4,i})$	$\boldsymbol{\mu}_{4,1}=[0\quad -4]$, $\boldsymbol{\Sigma}_{4,1}=\mathbf{I}_{2\times 2}$ $\boldsymbol{\mu}_{4,2}=[-4\quad 0]$, $\boldsymbol{\Sigma}_{4,2}=2\mathbf{I}_{2\times 2}$ $\boldsymbol{\mu}_{4,3}=[4\quad 0]$, $\boldsymbol{\Sigma}_{4,3}=3\mathbf{I}_{2\times 2}$

Note: $\boldsymbol{\alpha}1=\boldsymbol{\alpha}_2=\boldsymbol{\alpha}_3=\boldsymbol{\alpha}_4=[0.6\quad 0.3\quad 0.1]$.

TABLE III

THE CLASSIFICATION PERFORMANCE FOR GAUSSIAN DISTRIBUTION WITH HYBRID CLASSIFICATION METHODS COMPARED TO PURELY NON-PARAMETRIC CLASSIFIERS.

Hybrid neural network					Pure neural network				
	C_1	C_2	C_3	C_4		C_1	C_2	C_3	C_4
C_1	352	68	6	64	C_1	348	72	6	64
C_2	66	384	42	8	C_2	78	372	42	8
C_3	4	76	381	38	C_3	1	79	381	38
C_4	37	13	71	390	C_4	33	17	71	390
$\overline{P_{oc}}$	**75.35%**				$\overline{P_{oc}}$	**74.55%**			
Hybrid nearest neighbor					Pure nearest neighbor				
	C_1	C_2	C_3	C_4		C_1	C_2	C_3	C_4
C_1	340	63	18	69	C_1	334	69	18	69
C_2	58	352	74	16	C_2	78	332	74	16
C_3	7	95	315	82	C_3	15	87	315	82
C_4	70	11	65	365	C_4	69	12	65	365
$\overline{P_{oc}}$	**68.60%**				$\overline{P_{oc}}$	**67.30%**			

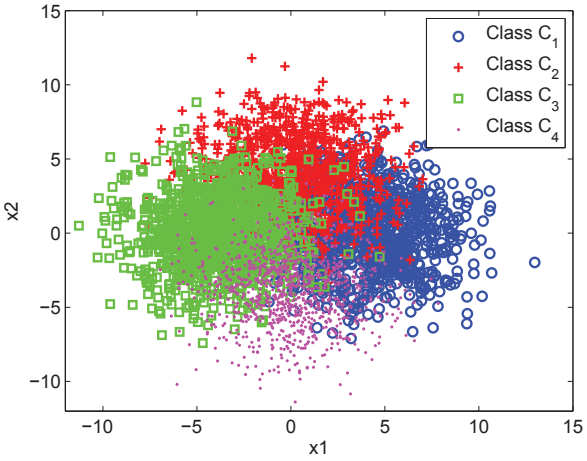

Fig. 2. A data distribution of four classes

In this simulation, we consider 1000 data samples for each class, and we use half of them as training data and the other half as test data. To evaluate the performance, we examine the classification results between the hybrid classification method with three purely non-parametric classifiers: nearest neighbor and neural network. The detailed comparison of classification performance is shown in Table. III where $\overline{P_{oc}}$ is the average of classification accuracy. It shows that, classification accuracy of hybrid classifiers is better than purely non-parametric classifiers and is close to the optimal Bayesian classifier (75.85%) where all class's distributions are completely known beforehand.

B. Power Quality Disturbance

We further apply the hybrid classification method for an important issue of power delivery and power industry: power quality (PQ) disturbance classification. Poor PQ may cause electricity blackouts, equipment failures or malfunctions, and financial loss. A quick identification of PQ disturbance could help to make a control decision which may avoid the subsequent influence. A study conducted by Lawrence Berkeley National Laboratory estimates that electric power outages and blackouts cost the U.S. about $80 billion annually [10]. The PQ disturbance signal is characterized by parameters that express amplitude swell or sag, harmonic pollution, reactive power, load unbalance, among others. This characterization of PQ disturbance indicates that the underlying data distribution for some classes may be modeled, while the others may be not. However, most existing power quality disturbances classifiers are based on the features selected from the raw data for which some obtainable class's distributions are disregarded, such as Self Organizing Learning Array (SOLAR) system based on wavelet transformation [11], inductive inference approach [12], SVM classification, etc. In contrast to these non-parametric classifiers, a Bayesian classifier can take advantage of the PQ disturbance model and use the raw measurements directly. Because of the use of all distribution models, the proposed Bayesian classifier is a generative classier with the analytic form of posterior probability for all classes without suffering from "curse of dimensionality". However, in practice, we have to mention that some distribution models are unknown, hence the purely Bayesian classifier wouldn't work.

In our simulation, we consider the same PQ disturbance models with seven different classes ($C1-C7$) shown in Table IV, which includes normal, swell, sag, harmonic, outage, sag with harmonic, swell with harmonic [11] [12]. We assume the distribution models of classes C_1, C_2, C_3 and C_4 are known, while the distribution models of remaining classes C_5, C_6 and C_7 are unknown. We compare the hybrid classifiers with these non-parametric classifiers including SOLAR and SVM. The SOLAR classification method is a self organization learning array system based on wavelet transformation. In [11], the classification performances based on SVM are also reported. In this simulation, we directly combine the SOLAR method and SVM into the hybrid classifier. The hybrid classification method with SOLAR based on wavelet transformation obtains 96.84% and the purely SOLAR method has 94.93%, while the hybrid classification method with C-

TABLE IV

THE CATEGORIES OF POWER QUALITY DISTURBANCE.

PQ Disturbance Type	Class Symbol	Signal Model	Parameters
Normal	C_1	$s_1 = A\sin(\omega_0 t)$	N/A
Swell	C_2	$s_2 = A(1+\alpha(u(t-t_1)-u(t-t_2)))\sin(\omega_0 t)$ $u(t) = \begin{cases} 1 & \text{if } t \geq 0 \\ 0 & \text{otherelse} \end{cases}$	$0.1 \leq \alpha \leq 0.8,\ T \leq t_2 - t_1 \leq 9T$
Sag	C_3	$s_3 = A(1-\alpha(u(t-t_1)-u(t-t_2)))\sin(\omega_0 t)$	$0.1 \leq \alpha \leq 0.8, T \leq t_2 - t_1 \leq 9T$
Outrage	C_4	$s_4 = A(1-\alpha(u(t-t_1)-u(t-t_2)))\sin(\omega_0 t)$	$0.9 \leq \alpha \leq 1, T \leq t_2 - t_1 \leq 9T$
Harmonic	C_5	$s_5 = A(\alpha_1\sin(\omega_0 t) + \alpha_3\sin(3\omega_0 t) + \alpha_5\sin(5\omega_0 t) + \alpha_7\sin(7\omega_0 t))$	$0.05 \leq \alpha_3 \leq 0.15, 0.05 \leq \alpha_5 \leq 0.15$ $0.05 \leq \alpha_7 \leq 0.15, \sum \alpha_i^2 = 1$
Sag with Harmonic	C_6	$s_6 = A(1-\alpha(u(t-t_1)-u(t-t_2)))$ $(\alpha_1\sin(\omega_0 t) + \alpha_3\sin(3\omega_0 t) + \alpha_5\sin(5\omega_0 t))$	$0.1 \leq \alpha \leq 0.9, T \leq t_2 - t_1 \leq 9T$ $0.05 \leq \alpha_3 \leq 0.15, 0.05 \leq \alpha_5 \leq 0.15, \sum \alpha_i^2 = 1$
Swell with Harmonic	C_7	$s_7 = A(1+\alpha(u(t-t_1)-u(t-t_2)))$ $(\alpha_1\sin(\omega_0 t) + \alpha_3\sin(3\omega_0 t) + \alpha_5\sin(5\omega_0 t))$	$0.1 \leq \alpha \leq 0.9, T \leq t_2 - t_1 \leq 9T$ $0.05 \leq \alpha_3 \leq 0.15, 0.05 \leq \alpha_5 \leq 0.15, \sum \alpha_i^2 = 1$

A: the amplitude of sine
ω_0: the angular frequency of sine
$u(t)$: the step function

SVM obtains 96.80% and the purely C-SVM gets 94.89%. It shows that the hybrid classifiers outperform the purely non-parametric classifiers for power quality disturbance classification.

V. CONCLUSION AND FUTURE WORKS

In this paper, a novel classification framework is proposed to address a new classification problem where only the distribution model of some classes are known. Instead of disregarding these important information directly in non-parametric classifiers, the proposed hybrid classification framework combines the Bayesian classifier and non-parametric classifier to make classification. It makes the best use of the knowledge of known distribution models to improve classification performance. Theoretically proofs and experimental results show that the proposed hybrid classifier has a better performance than these purely non-parametric classifiers for the data with partial models. Currently, we demonstrate the effectiveness of proposed hybrid classification method in the partially known models. We plan to apply the hybrid classification method into the real-life classification problems with partially known models in which only the non-parametric classification methods are employed. With the knowledge of known distributions of some classes, we expect that we can obtain much better classification performance.

REFERENCES

[1] S.Kay, *Fundamentals of Statistical Signal Processing: Detection Theory*. NJ: Prentice-Hall: Englewood Cliffs, 1998.

[2] C. E. Rasmussen, "Gaussian processes for machine learning," 2006.

[3] A. Jordan, "On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes," *Advances in neural information processing systems*, vol. 14, p. 841, 2002.

[4] S. Tong and D. Koller, "Restricted bayes optimal classifiers," in *AAAI/IAAI*, pp. 658–664, 2000.

[5] R. Raina, Y. Shen, A. Mccallum, and A. Y. Ng, "Classification with hybrid generative/discriminative models," in *Advances in neural information processing systems*, p. None, 2003.

[6] A. Fujino, N. Ueda, and K. Saito, "A hybrid generative/discriminative approach to semi-supervised classifier design," in *PROCEEDINGS OF THE NATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE*, vol. 20, p. 764, Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2005.

[7] Y. Li, L. Shapiro, and J. A. Bilmes, "A generative/discriminative learning algorithm for image classification," in *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, vol. 2, pp. 1605–1612, IEEE, 2005.

[8] T. Roos, H. Wettig, P. Grünwald, P. Myllymäki, and H. Tirri, "On discriminative bayesian network classifiers and logistic regression," *Machine Learning*, vol. 59, no. 3, pp. 267–296, 2005.

[9] T. Van Kasteren, G. Englebienne, and B. J. Kröse, "An activity monitoring system for elderly care using generative and discriminative models," *Personal and ubiquitous computing*, vol. 14, no. 6, pp. 489–498, 2010.

[10] K. H. LaCommare and J. H. Eto, "Understanding the cost of power interruptions to us electricity consumers," 2004.

[11] H. He and J. A. Starzyk, "A self-organizing learning array system for power quality classification based on wavelet transform," *Power Delivery, IEEE Transactions on*, vol. 21, no. 1, pp. 286–295, 2006.

[12] T. K. A. Galil, M. Kamel, A. M. Youssef, E. F. E. Saadany, and M. M. A. Salama, "Power quality disturbance classification using the inductive interference approach," *IEEE Trans. Power Del.*, pp. 1812–1818, Oct. 2006.