PROBABILITY CALIBRATION BY THE MINIMUM AND MAXIMUM PROBABILITY SCORES IN ONE-CLASS BAYES LEARNING FOR ANOMALY DETECTION

GUICHONG LI[1], NATHALIE JAPKOWICZ[1], IAN HOFFMAN[2], R. KURT UNGAR[2]

ABSTRACT. One-class Bayes learning such as one-class Naïve Bayes and one-class Bayesian Network employs Bayes learning to build a classifier on the positive class only for discriminating the positive class and the negative class. It has been applied to anomaly detection for identifying abnormal behaviors that deviate from normal behaviors. Because one-class Bayes classifiers can produce probability score, which can be used for defining anomaly score for anomaly detection, they are preferable in many practical applications as compared with other one-class learning techniques. However, previously proposed one-class Bayes classifiers might suffer from poor probability estimation when the negative training examples are unavailable. In this paper, we propose a new method to improve the probability estimation. The improved one-class Bayes classifiers can exhibits high performance as compared with previously proposed one-class Bayes classifiers according to our empirical results.

1. INTRODUCTION

One-class classification [9][22][23] is a technique that builds a classifier on the positive class only by learning the data characteristics and building the decision boundary to discriminate the positive class and the negative class. In general, this is achieved by deriving the induction algorithm from the corresponding supervised learning algorithm. For example, one-class Support Vector Machine (OCSVM) [22], which is derived in the way similar to that of the corresponding supervised SVM, learns the maximum margin between the positive examples and the origin.

Unlike OCSVM one-class Bayes classification applies Bayes learning to build one-class classifiers. For example, one-class Naïve Bayes, which is derived from the corresponding supervised Naïve Bayes, builds one-class classifier by assuming conditional independences among attributes given the class. One-class Bayesian Network, which is derived from the corresponding supervised Bayesian Network, builds a Bayesian Network on the positive class only by learning dependencies of attributes from the positive class.

One-class Bayes classification has been widely used for anomaly detection [3][19], e.g., network intrusion detection [7], disease outbreak [28], wireless sensor detecting [20], spam filtering [26], etc. The salient advantage is that using Bayes' rule it can produce probability scores, which can be used for defining anomaly score as the degree in which a test example is detected to be an abnormal case for anomaly detection.

The main issue is that previously proposed one-class Bayes learning techniques suffer from some limitations to perform probability estimation properly. For example, a simple one-class Naïve Bayes [25] directly applies the supervised Naïve Bayes to the positive class with the assumption that there is at least one negative case to estimate conditional probability given the negative class in nominal cases for one-class learning. There are at least three limitations behind this assumption: first, it is ineffective when an application is involved with continuous variables; secondly, the assumption suffers from the curse of dimensionality because it is insufficient in high dimension; thirdly, the method is unreliable in one-class learning when it is dependent of the assumption about the

[1] Computer Science of University of Ottawa, {jli136, nat}@site.uottawa.ca.com
[2] Radiation Protection Bureau, Health Canada, {ian.hoffman, kurt.ungar}@hc-sc.gc.ca

negative class distribution. For another example, Naïve Bayes Positive Class [9], which is an early proposed one-class Naïve Bayes, only performs classification without outputting class membership probability.

Similarly, in previous research [8][28], one-class Bayesian Network, which is built on the positive class by using the corresponding supervised discrete Bayesian Network, produces probability scores, which are not straightforward to be a proper class membership probability. As a result, one-class Bayes classifiers often suffer from poor performance for anomaly detection in complex applications. These limitations unexpectedly degrade the performance of one-class Bayes learning in many circumstances where probability estimation becomes crucial when the costs of false positive cases and false negative cases are different [13].

Although people have proposed some approaches for probability calibration in decision trees and Naïve Bayes [30]. However, these methods such as the binning method [30], which is associated with negative examples for Naïve Bayes, are inapplicable because there are positive training examples only in one-class learning.

In this paper, our main work is to propose a new method to improve one-class Bayes learning algorithms such that they can produce class membership probability properly. The main advantage is that it is independent of the negative class distribution for one-class learning. It is more effective than previously proposed methods in practical applications consisting of either nominal or continuous variables. The improved one-class Bayes learning algorithms are compared with previously proposed one-class Bayes learning algorithms by conducting experiments on the benchmark datasets from the UCI repository [17] and two practical applications for justification.

2. PRELIMINARY

2.1 One-Class Learning and Anomaly Detection

The basic definition of one-class classification [23], also called single class learning [9][22], has been described in various works.

One-Class Learning (OCL) is essentially a two-class classification task which follows an underlying binary distribution. A One-Class (OC) classifier is built on the single known class to predict a new pattern as being a member of the known class or not. If it is not predicted to be a member of the known class, then it is automatically assumed to belong to the unknown class whose distribution is different from that of the known class.

The single known class is also called the *positive* class or the *target*, *normal* class while the unknown class to be estimated is called the *negative* class or the *outlier*, *novelty* [21], *anomaly* class [4] in different applications.

Anomaly detection uses techniques to find patterns in data that do not conform to expected behavior [3][4]. The goal can be achieved by producing an anomaly score [4], also called outlier factor [2] or outlying degree [31], which is the degree to which an instance belongs to an anomaly class. Given an instance x, the decision rule using the anomaly score for predicting its class label y is defined as

$$y(x) = \begin{cases} 0, \textit{positive}, \textit{if } \textit{AnomalyScore}(x) < s_0, \\ 1, \textit{negative}, \text{otherwise} \end{cases} \tag{2.1}$$

where s_0 is the cutoff value of the anomaly score.

According to whether labeled data and unlabeled data are available in the training set, anomaly detection techniques consist of three categories: unsupervised learning,

supervised learning, and semi-supervised learning [19]. Semi-supervised learning applies to positive cases and an abundant unlabeled database. There is, however, an extreme case in which people can obtain as many reliable positive cases as they want while obtaining negative cases is impractical. Unlabeled data in such settings are more likely to be positive cases. Obtaining negative cases is, then, prohibitively expensive. In general, in such cases, a few labeled negative examples or artificial negative examples are what are used for validating the false negative rate during training [23]. This is the essential distinction between one-class learning (the latter) and semi-supervised learning (the former).

Our recent research has been focused on the application of machine learning techniques to detect nuclear emissions from medical isotope production facilities. The task consists of classifying spectra obtained from NaI scintillation detectors located at two different locations in the Ottawa valley. Medical isotope production at Chalk River Laboratories routinely results in emissions of various radioactive isotopes that can easily be observed in the 15 minute sample acquisition intervals of the NaI detectors. The task is to classify each spectra as having nuclear emissions present or not in the presence of a fluctuating background. The task is made more difficult in that spectra acquired during precipitation events dramatically alter the spectra from those typical of normal background and of emission events. For general environmental radiation monitoring the observations of the negative class, or spectra containing nuclear emissions superimposed on a natural background environment are difficult to obtain, while the observations of the positive class for normal background are common. Insufficient sampling of the negative class may not describe the underlying distribution properly and a model that relied on such data might lead to a failure to predict the abnormal environmental changes. In particular, labeling a sufficient number of abnormal cases can be unreliable and unrealistic. One-class learning techniques in machine learning are, therefore, necessary, for this type of environmental radiation monitoring.

Empirically, two-class supervised learning is superior to one-class learning when the positive class and the negative class are properly defined [18][23][29]. One-class learning, also called Negative selection [32], can be harder than two-class learning due to higher sample complexity [23].

2.2 Bayes Learning

Given a training set with a probability distribution P, in supervised learning, Bayesian learning defines a classifier with a minimized error, i.e.,

$$\begin{aligned} y_i = c_i = \arg\max_{c_i \in C} P(c_i \mid x) = \arg\max_{c_i \in C} P(x,c_i)/P(x) \equiv \arg\max_{c_i \in C} P(x|c_i)P(c_i) \\ = \arg\max_{c_i \in C} P(a_1, a_2, ..., a_n \mid c_i)P(c_i) \end{aligned} \tag{2.2}$$

Naïve bayes (NB) [10] assumes the probabilities of attributes a_1, a_2, ..., a_n to be conditionally independent given the class c_i. Therefore, $P(x|c_i)$ from the right side of (2.2) becomes

$$P(x \mid c_i) = P(a_1, a_2, ..., a_n \mid c_i) = \prod_{j=1}^{n} P(a_j \mid c_i) \tag{2.3}$$

For discrete attribute a_j, $P(a_j \mid c_i)$ can be estimated by using Maximum Likelihood Estimation (MLE) with Laplace smooth, i.e.,

$$\hat{P}(a_{jk} \mid c) = \frac{n_{jkc} + 1}{n_c + l} \tag{2.4}$$

where n_{jkc} is the number of occurrences of the attribute value a_{jk} in the class c, and n_c is the number of examples in the class c, and l is the number of distinct values in the attribute a_j.

Smoothing in (2.4) assumes that each attribute value at least occurs one time in each class by following a Dirichlet prior distribution over a_j. In particular, in one-class learning, it can avoid the result of zero for probability estimation when the negative class $c = c_1$ is empty if the traditional supervised Naïve Bayes algorithm is used. In the same way, the prior probability for the negative class c_1 is estimated by $\hat{P}(c_1) = 1/(m+2)$, where m is the total number of training examples.

For continuous attributes a_j, $P(a_j \mid c_i)$ can be estimated by using Gaussian Estimator (GE) or Parzen–window density estimator. The latter is defined as

$$P(x) = \frac{1}{n}\sum_{j=1}^{n}\frac{1}{h_n^d}K\left(\frac{x - x_j}{h_n}\right) \tag{2.5}$$

where $K(x)$ is a kernel function placed at each observation x_j in the d-dimension feature space in the window with the width h_n. However, the estimation of the parameters related to the negative class c_1 in one-class learning becomes impossible if the traditional Naïve Bayes algorithm is used. To avoid this, $P(a_j \mid c_1)$ can be assigned by a small real number default for the computation of the product in (2.3).

A Bayesian Network (BN) [10][15] with a directed acyclic graph (DAG) describes a joint probability distribution P on a set of random variables $X = \{x_j\}$, $j = 1, \ldots, n$, by encoding independencies among variables X given their parents. Further, BN is used for classification by estimating the conditional probability $p(x|c_i)$ and $P(c_i)$ in (2.2). Given an observation $x = (a_1, \ldots, a_n)$, $p(x|c_i)$ can be rewritten as

$$P(x \mid c_i) = P(a_1, a_2, \ldots, a_n \mid c_i) = \prod_{j=1}^{n} P(a_j \mid a_{j+1}, \ldots, a_n, , c_i) \tag{2.6}$$

According to the independence assumptions encoded in DAG, (2.6) can be rewritten as

$$P(x \mid c_i) = P(a_1, a_2, \ldots, a_n \mid c_i) = \prod_{j=1}^{n} P(a_j \mid \pi_j, c_i) \tag{2.7}$$

where $p(a_j|\pi_j, c_i)$ (i.e., $x_j = a_j$) is a class conditional probability that represents that x_j is independent of nonparent nodes given its parent variables π_j and the class c_i.

In practice, the DAG G can be learned by using a hill climbing algorithm to search for the dependent relationships among variables. Because the optimized DAG is intractable [5], the hill climbing with a restricted order of variables is usually applied to build DAG [6]. A more efficient technique for building DAG is to find a maximal weighted span tree [15].

3. RELATED WORK

Bayesian Learning such as Naïve Bayes and Bayesian Network has been used for one-class learning [9][20][28]. The main idea is that a Bayes classifier can produce the probability score of a given input for the positive class. Given a threshold, the input belongs to the positive class if the estimated probability of the input is higher than the threshold. Otherwise, it is regarded as a negative case.

We introduce two kinds of one-class Bayes learning: one-class Naïve Bayes such as Naïve Bayes Postive Class (NBPC) [9], and One-Class Bayesian Network (OCBN)[28]. They are derived from the corresponding Bayes classifiers for one-class learning.

3.1 One-class Naïve Bayes

Naïve Bayes Positive Class (NBPC) algorithm is a one-class Naïve Bayes method, which is derived from the original supervised Naïve Bayes algorithm [9]. Notice that we can only estimate the prior probability of the positive class because the negative class is not in the training set. Using the traditional Naïve Bayes inductive algorithm, the prior probability $P(c_0)$, defined in (2.2), of the positive class [9] is estimated as a fraction close to 1 by assuming at least one negative case for Laplace smooth. Further, because only positive cases are available in the training set, conditional probabilities of nominal attributes given the negative class can be estimated by assuming at least one negative case for Laplace smooth, as described in (2.4).

During the training period, the parameters of NBPC are calculated as in the traditional Naïve Bayes except an additional parameter, which is called the *target rejection threshold* τ, and is calculated in (3.1). For testing, a new instance is identified as positive if the probability output by the NBPC is greater than or equal to τ. Otherwise, it is a negative case.

$$\tau = \min\{p(c_0 | x^{(k)})\} = \min\{p(c_0) \times \prod_{j=1}^{n} p(a_j^{(k)} | c_0)\} \quad (3.1)$$

where $x^{(k)} = (a_1^{(k)}, ..., a_n^{(k)}) \in D_m^n$, $k = 1, ..., m$, and D_m^n is a training set with n attributes and m training examples. Therefore, NBPC does not produce anomaly score but classification.

A simple one-class Naïve Bayes [25], which is also called the simple OCNB, actually is a Naïve Bayes built on the positive class only. It is similar to NBPC except the target rejection threshold. That is, it performs Laplace smooth by assuming at least one case to estimate the prior probability $p(c_i)$ and conditional probability $p(a_j \mid c_i)$ for nominal attributes, as discussed in Section 2.2.

3.2 One Class Bayesian Network

A Bayesian Network (BN) [6] is a probability model that represents a joint probability distribution with a direct graph. The graphical structure describes the conditional dependences among attributes while it also encodes the conditional independences of the attributes. It can describe complex relationships between attributes instead of using the conditional independence assumption of one-class Naïve Bayes.

Discrete Bayesian Networks have been used for anomaly detection in the multi-class setting [8][28]. This corresponding algorithm for one-class learning is called one-class Bayesian Network (OCBN), which is expected to be better than OCNB in some complex learning tasks because it can learn the dependencies of attributes.

During training, the Bayesian Network structure in the OCBN can also be built by using a hill climbing algorithm with a restricted order of variables [6] as in the original BN; the parameters for the conditional probability tables (CPTs) related to the negative class is initialized by using Laplace smooth as in NB by assuming that one nominal attribute value at least happen one time in training examples. For testing, the decision rule defined in (2.2) is used for predicting the test example. As we can see, this one-class Bayesian Network is also called the simple OCBN similar to the simple OCNB because it is just a BN built on the positive class only.

As we can see, both the simple OCNB and the simple OCBN are dependent of the negative class due to their assumptions about the negative distribution while NBPC does not perform probability estimation. They are only applicable for nominal cases.

Further, in previous research, to improve the probability estimation in Naïve Bayes for supervised learning, the binning method [30] is first to sort training examples according to probability scores and dividing the sorted set into 10 bins with the lower and upper boundary during the training time. For testing, a new example x is placed in a bin b according to its score. The corrected probability $P(c_i|x) = n_i' / n'$, where n' is the number of training examples in b; n_i' is the number of training examples that actually belongs to the class i in b. However, the binning method is inapplicable in one-class Bayes learning because there are only positive examples for training.

4. PROBABILITY ESTIMATION AND ANOMALY SCORE

Although classification is required, the probability estimation of the class membership of a new instance is more critical in some circumstances. In particular, if the costs of misclassifications for the false positive and false negative cases are different, the probability estimation helps Cost-Sensitive learning [11][13][30]. This is often true when applying one-class learning to many practical applications.

In general, an anomaly detection technique always outputs the anomaly score for decision, as defined in (2.1). If the anomaly score falls within [0, 1], it can be easily transformed into the class membership probability by defining $p(c_1|x)$ = AnomalyScore(x) and $p(c_0|x) = 1 - p(c_1|x)$. Both can be mutually exchanged, and can be directly used for classification.

The main issue is that some previously proposed one-class Bayes algorithms do not perform probability estimation properly. For example, in NBPC, although the decision rule is defined according to τ in (3.1), the estimated probability $P(c_0|x)$ in (2.2) is not regarded as a proper class membership probability while it becomes a probability score. Because the negative class is unavailable in the training set, the prior class probability $P(c_i)$ and the marginal prior probability $P(x)$ in (2.2) cannot be estimated properly from the data. Note that in supervised learning the marginal prior probability $P(x)$ is omitted.

The probability estimation for class membership is not straightforward from (2.2) when negative training examples are unavailable. In the simple one-class Naïve Bayes, as discussed in Section 3.1, the assumption that there is at least one negative case for the probability estimation is unreliable in practice. As a result, it is not expected that the simple one-class Naïve Bayes performs probability estimation properly. No anomaly score is expected in these one-class Bayes approaches for anomaly detection.

When the minimum probability score in (3.1) is defined as the cutoff τ for decision, we also can obtain the maximum probability score $\hat{\tau} = \max\{p(c_0|x^{(k)})\}$, $k = 1, \ldots, m$. As a result, we can define a new method for probability estimation in one-class Naïve Bayes, e.g., NBPC, according to τ and $\hat{\tau}$, in (4.1).

$$\hat{p}(c_0 \mid x) = \begin{cases} 0.5 + 0.5 \times \left(p(c_0 \mid x) - \tau\right) / \left(\hat{\tau} - \tau + \varepsilon\right), \tau \le p(c_0 \mid x) \le 1 \\ 0.5 + 0.5 \times \left(p(c_0 \mid x) - \tau\right) / \tau, \text{otherwise} \end{cases} \tag{4.1}$$

where a sufficiently small number, e.g., ε = 0.001, is given; and $p(c_0|x)$ is a probability score; $\hat{p}(c_0 \mid x)$ is the resulting class membership probability for the positive class, and $\hat{p}(c_0 \mid x) = 0.5 + \varepsilon$ if $p(c_0|x) = \tau$; $\hat{p}(c_1 \mid x) = 1 - \hat{p}(c_0 \mid x)$, that is, $0 \le \hat{p}(c_i \mid x) \le 1$, and the sum is equal to 1. In general, τ is nonzero and $\hat{\tau} > \tau$. To avoid an invalid denominator due to $\hat{\tau} = \tau$, the denominator is added with ε. This extreme case also means that the classifier performs poor probability estimation. As we can see, $\hat{p}(c_0 \mid x)$ is monotonic increasing with the probability score $p(c_0|x)$.

The minimum probability score τ and the maximum probability score $\hat{\tau}$ are useful for probability calibration because one cannot expect that the probability scores $p(c_0|x)$ fully spread over the interval [0, 1]. $\hat{p}(c_0 \mid x)$, defined in (4.1), is a probability function with respect to the probability score $p(c_0|x)$ and two related parameters, $\hat{\tau}$ and τ, i.e., $\hat{p}(c_0 \mid x) = f(p(c_0|x), \hat{\tau}, \tau)$. $\hat{p}(c_i \mid x)$, i = 0, 1, can be properly used as class membership probabilities. Similarly, the probability estimation method in OCBN can be defined as in (4.1).

It can be easily seen that the probability function, defined in (4.1), is independent of the negative class distribution. This property is more important when negative examples are unavailable because they are too prohibitively expensive to obtain in some cases. The critical issue is that τ, as defined in (3.1), might be inappropriate for target rejection in noise circumstances. In one-class learning, the *target rejection rate r* is defined as the proportion of training examples that will be classified as the negative class. Therefore, (3.1) can be rewritten as

$$\tau = \min(p(c_0|x^{(k)}), r \times m), k = 1, \ldots, m \tag{4.2}$$

where min(P, l) function returns the lth minimum value of P.

```
OneClassNaiveBayes algorithm
Input D: training set
      r: target rejection rate
Output OCBN: OneClassNaiveBayes
   classifier
1  assuming c0: target class, c1:
   the negative class
2  calculate p(ak|c0), p(c0),
   where k = 0,…,l-1; l: the
   number of attribute; MLE and GE
   for nominal and continuous
   attributes
3  τ = min(p(c0|xi, r×m) in (4.2)
4  τ̂ = max{p(c0|xi}, i = 0,…,m-1
5  return OCNB(p(aj|c0), p(c0), τ,
   τ̂), j = 0,…, k-1,
end OCNB
Proc test(x)
6   get p(x|c0), p(c0) from
   p(ak|c0) in OCNB
7   calculate p̂(c0|x), p̂(c1|x)=1-
    p̂(c0|x), according to (4.1)
8   return cj = arg max_j p̂(cj|x), j=0,1
end test
```

```
OneClassBayesNet algorithm
Input D: training set
      r: target rejection rate
Output OCBN: OneClassNaiveBayes
  classifier
1  assuming c0: target class, c1:
   the negative class
2  learning Bayesian Nework
      structure
3  calculate p(ak|Pk,c0), p(c0),
   where k = 0,…,l-1; l: the
   number of attribute; Pk is the
   parents of ak
4  τ = min(p(c0|xi, r×m) in (4.2)
5  τ̂ = max{p(c0|xi)}, i = 0,…,m-1
6  return OCBN(p(ak|Pk,c0), p(c0),
   τ, τ̂), k = 0,…, l-1,
end OCBN
Proc test(x)
7  get p(x|c0), p(c0) from p(ak|Pk,
   c0) in OCBN
8   calculate p̂(c0|x), p̂(c1|x)=1-
    p̂(c0|x), according to (4.1)
9   return cj = arg max_j p̂(cj|x), j=0,1
end test
```

5. IMPROVED METHOD

According to the above discussion, we propose OneClassNaiveBayes (OCBN) and OneClassBayesNet (OCBN) algorithms, which improve previously proposed one-class Naïve Bayes and one-class Bayesian Network algorithms, respectively. The algorithms are derived from the traditional Naive Bayes and Bayesian Network. During the training time, the most parameters are calculated in the same way as in the original supervised methods

except two additional parameters described as above. OCBN and OCBN will learn two additional parameters: τ and $\hat{\tau}$, as defined in Steps 3, 4 of the OCNB algorithm and Steps 4, 5 of the OCBN algorithm, and use the proposed method for probability estimation in their test procedures.

As in the original Naive Bayes, the parameters of OCNB can be calculated by Gaussian estimator or Parzen-window density estimator for continuous attributes. It can be also built by discretizing continuous attributes. Further, as in the original BN, OCBN can be also built by a hill climbing algorithm with a restricted order on attributes for searching its network structure [6], or by learning a maximum weight span tree for the structure [15].

The main concern is that the discretization cannot be achieved by using the supervised method based on entropy [14] because no negative examples are available. Therefore, the 10-bined unsupervised method is used for discretization in the discrete OCBN.

6. EXPERIMENTS

6.1 Datasets

We chose 30 benchmark datasets from the UCI repository [17], and two real datasets: Ozone Level Detection [1][33] and OttawaRPB for ozone level detection and the environment radiation monitoring, respectively. Because the benchmark datasets have been built in high quality for supervised learning, and they often contain continuous and nominal attributes, this provides us to evaluate the new method for one-class Bayes learning on various domains. The characteristics of all datasets are described in Table 1.

Ozone Level Detection datasets (the eight hour peak set and one hour peak set) were collected from 1998 to 2004 at the Houston, Galveston, and Brazoria area. One hour peak set (Ozone in Table 1) is chosen by ignoring the date in our experiment. In the dataset, the 72 continuous attributes contains various measures of air pollutant and meteorological information for detecting ozone days. There are 73 ozone days labeled as the negative class in the class attribute in the dataset while the majority class consists of positive examples.

The OttawaRPB for the environmental radiation monitoring data is a complex domain consisting of 512 continuous attributes, the class attribute, and 2914 labeled instances with only 129 negative examples. OttawaRPB is described in Section 2.

For experiments on the benchmark datasets, each dataset was transformed into a binary domain consisting of the majority class and the rest of the data in advance of training time.

All missing values were replaced with their modes and means for nominal attributes and continuous attributes, respectively, by using the unsupervised ReplaceMissingValues method in Weka [27] ahead of training. During training, each one-class classifier is built on only the majority class as the positive class (target class) of the binary domain. The majority class in the binary domain might be different from that one in the original dataset. Therefore, the positive class is always larger than the negative class, as shown in Table 1. As we can see, they are generally class imbalanced. The largest ratio of the positive class to the negative class is 33.74:1 in the Ozone case.

6.2 Algorithms for comparison

We used the Weka data mining and machine learning package [27] to implement two one-class Bayes algorithms: one-class Naïve Bayes (OCNB) and one-class Bayesian Network

(OCBN) by improving the previous one-class Naïve Bayes approaches such as NBPC and the simple one-class NB for probability estimation. The improved OCNBs and improved OCBNs can be adapted with various settings for OCL, as described in Table 2.

Table 1. Datasets in our experiments. The 30 benchmark datasets from the UCI repository, and two real datasets: ozone level detection (Ozone) and ottawaRPB for practical applications. #maj: the size of the majority class in the original dataset; #pos is the size of the majority class in the binary class; the ratio is given by #pos / (#ins-#pos).

Datasets	#attr	#ins	#c	#maj	#pos	ratio	Datasets	#attr	#ins	#c	#maj	#pos	ratio
Anneal	39	898	6	684	684	3.20	Letter	17	20000	26	813	19187	23.60
Audiology	70	226	24	57	169	2.96	Lymph	19	148	4	81	81	1.21
Autos	26	205	6	67	138	2.06	Mushroom	23	8124	2	4208	4208	1.07
Balance-s	5	625	3	288	337	1.17	P-tumor	18	339	21	84	255	3.04
Breast-w	10	699	2	458	458	1.90	Segment	20	2310	7	330	1980	6.00
Colic	23	368	2	232	232	1.71	Sick	30	3772	2	3541	3541	15.33
Credit-a	16	690	2	383	383	1.25	Sonar	61	208	2	111	111	1.14
Diabetes	9	768	2	500	500	1.87	Soybean	36	683	18	92	591	6.42
Glass	10	214	6	76	138	1.82	Splice	62	3190	3	1655	1655	1.08
Heart-s	14	270	2	150	150	1.25	Vehicle	19	846	4	218	628	2.88
Hepatitis	20	155	2	123	123	3.84	Vote	17	435	2	267	267	1.59
Hypothyroid	30	3772	4	3481	3481	11.96	Vowel	14	990	11	90	900	10.00
Ionosphere	35	351	2	225	225	1.79	Waveform	41	5000	3	1692	3308	1.96
Iris	5	150	3	50	100	2.00	Zoo	18	101	7	41	60	1.46
Kr-vs-kp	37	3196	2	1669	1669	1.09	Ozone	73	2536	2	2463	2463	33.74
Labor	17	57	2	37	37	1.85	OttawaRPB	513	2941	2	2812	2812	21.80

For example, OCNB-Parzen is an improved OCNB with the Parzen-window density estimator. OCNB-SimpleGaussian, OCNB-SimpleParzen, and OCNB-SimpleDiscretize are actually the traditional supervised Naïve Bayes classifiers directly built on the positive class only. They perform as the simple one-class Naïve Bayes with different settings. Note that the improved OCNB or the simple OCNB with different settings produces the same results on nominal domains. In the OCBN-K2, the Bayesian structure is learned by using a hill climbing algorithm with a restricted order of variables [6], and the conditional probability tables are directly estimated from data. Our purpose is to compare the improved OCNB and improved OCBN with simple Bayesian learning methods for one-class learning. Finally, we also show two the original Naïve Bayes and Bayesian Network for supervised learning on the two practical applications.

The most parameters in OCNB or OCBN are the same as those of NB or BN, respectively, except the target rejection rate (TRR). The improved OCNB and the improved OCBN need to adjust the TRR for training the related minimum probability score τ. On the other hand, OCBN and BN have two main parameters for training: the estimator for conditional probability tables (CPTs) and the search algorithm for the network structure. The simple estimator is chosen for estimating the CPTs directly from data while several typical search algorithms such as K2, Hill Climbing, and TAN [15] are set in our experiments, as described in Table 2.

For experiments over the 30 benchmark datasets with small feature space (≤ 70), OCNB and OCBN are set with a default for TRR = 0.0, i.e., all the positive examples are accepted as true positive cases. For experiments over the two large datasets with large feature space (> 70), we conducted experiments with different TRR settings for optimization.

Table 2. Algorithms used in experiments: One-class Naïve Bayes (OCNB), one-class Bayesian Network (OCBN) with various settings for one-class Bayes learning; Naïve Bayes (NB) and Bayesian Network (BN) with defaults in Weka for supervised learning.

Algorithms	Descriptions
OCNB-Gaussian	Improved one-class Naïve Bayes with Gaussian estimator
OCNB-Parzen	Improved one-class Naïve Bayes with Parzen-window density estimator
OCNB-Discretize	Improved one-class Naïve Bayes with discretization
OCNB-SimpleGaussian	Naïve Bayes with Gaussian Estimator for OCL
OCNB-SimpleParzen	Naïve Bayes with Parzen-window density estimator for OCL
OCNB-SimpleDiscretize	Naïve Bayes with discretization for OCL
OCBN-K2	Improved one-class Bayesian Network with a restricted order of variables
OCBN-Hill	Improved one-class Bayesian Network with Hill climbing search
OCBN-TAN	Improved one-class Bayesian network with TAN search
OCBN-SimpleK2	Bayesian Network with a restricted order of variables for OCL
OCBN-SimpleHill	Bayesian Network with Hill climbing search for OCL
OCBN-SimpleTAN	Bayesian Network with TAN search [15] for OCL
NB	Naïve Bayes with default Gaussian estimator
BN	Bayesian Network with default K2 search

6.3 Results

Our experiments were conducted by running 10 times the 10-cross validations. In each run, the dataset is separated into 10 fold by stratified sampling. In turn, one fold is held out for test, other folds are used for training. However, one-class classifiers were built on only the positive class in the training set while two-class classifiers were built on the whole training set containing the positive class and negative class. Therefore, the simple OCNB and the simple OCBN are built on the different portion of the training set as compared with the supervised NB and BN. The resulting classifiers were tested on the test set.

The area under ROC curve (AUC) [16] is used for evaluation in our experiments. The AUC of a classifier is equivalent to the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance [12][16]. The AUC's salient advantage is to evaluate performance without specifying a threshold. It has been suggested as the preferred metric rather than the misclassification rate to evaluate a model [12]. In our experiments, the AUCs obtained in the 10-cross validations are averaged for evaluation.

To evaluate the proposed method for probability estimation in one-class Bayes learning, we first analyze relative performance with respect to AUC between the improved OCNBs and the simple OCNBs. This can be done first by using the ratio of OCNB-SimpleDiscretize's AUC to OCNB-Parzen's AUC, as shown in Figure 1, where the diagonal line reflects the relative performance of OCNB-SimpleDiscretize against the compared algorithm; the vertical dotted line at x = 1.0 reflects the relative performance of OCNB-SimpleDiscretize against OCNB-Parzen; the horizontal dotted line at y = 1.0 reflects the relative performance of OCNB-Parzen against the compared algorithm.

The vertical dotted lines at x = 1.0 from (a) to (d) in Figure 1 only reflects the relative performance of OCNB-SimpleDiscretize against OCNB-Parzen. As we can see, OCNB-Parzen outperforms OCNB-SimpleDiscretize in most cases because most points are located at the left side of the vertical line. The horizontal dotted lines at y = 1.0 from (a) to (d) reflect the relative performance of OCNB-Parzen against the compared algorithm. As we can see, the OCNB-Parzen outperforms other OCNB in most cases because most points are below these horizontal lines. In particular, the improved OCNB-Parzen is much more successful than the OCNB-SimpleParzen for one-class Bayes learning over various

domains, as show in (*d*). In addition, the improved OCNB-Gaussian is better than the OCNB-SimpleGaussian on average according to (*a*) and (*c*) because more points are below the horizontal dotted line at *y* = 1.0 in (*c*) than in (*a*); according to (*b*), OCNB-Discretize is competitive with OCNB-SimpleDiscretize in most cases because most points lie on the diagonal line. In a word, the improved OCNB is more successful than simple OCNB for one-class learning over various domains, and the improved OCNB-Parzen is best among all OCNB.

Similarly, we show the relative performance between the improved OCBNs and simple OCBNs by using the ratio of OCBN-SimpleTAN's AUC to OCBN-TAN's AUC in Figure 2. From the vertical dotted lines at *x* = 1.0, OCBN-SimpleTAN are tied with OCBN-TAN in most cases because most points lie on the vertical lines. However, OCBN-TAN outperforms other improved OCBNs and other simple OCBNs because most points are below the horizontal lines at *y* = 1.0 from (*a*) to (*d*). In addition, because most points crossing those diagonal lines are located toward the right-bottom corner, this shows that the OCBN-SimpleTAN is superior to other OCBNs in most cases except the improved OCBN-TAN. In a word, the improved OCBN-TAN and OCBN-SimpleTAN are better than other OCBN while both are tied with each other in most cases.

Experimental results for the comparison between two one-class Bayes methods: OCNB and OCBN are shown in Figure 3 by their relative performances between OCBN-TAN and OCNB-Parzen, OCBN-TAN and other OCNB classifiers. It is easy to see that OCBN outperforms OCNB in most cases from the 30 benchmark datasets in the current settings.

We conducted the paired t-test for comparison between the improved one-class Bayes learning methods and all simply one-class Bayes learning methods. The results were summarized in Table 3. As a result, the improved OCBN-TAN seems not to exhibit super performance as compared with the simple OCBN-TAN while OCNB-Parzen and OCBN-TAN are better than other related simple one-class Bayes learning methods, as shown in

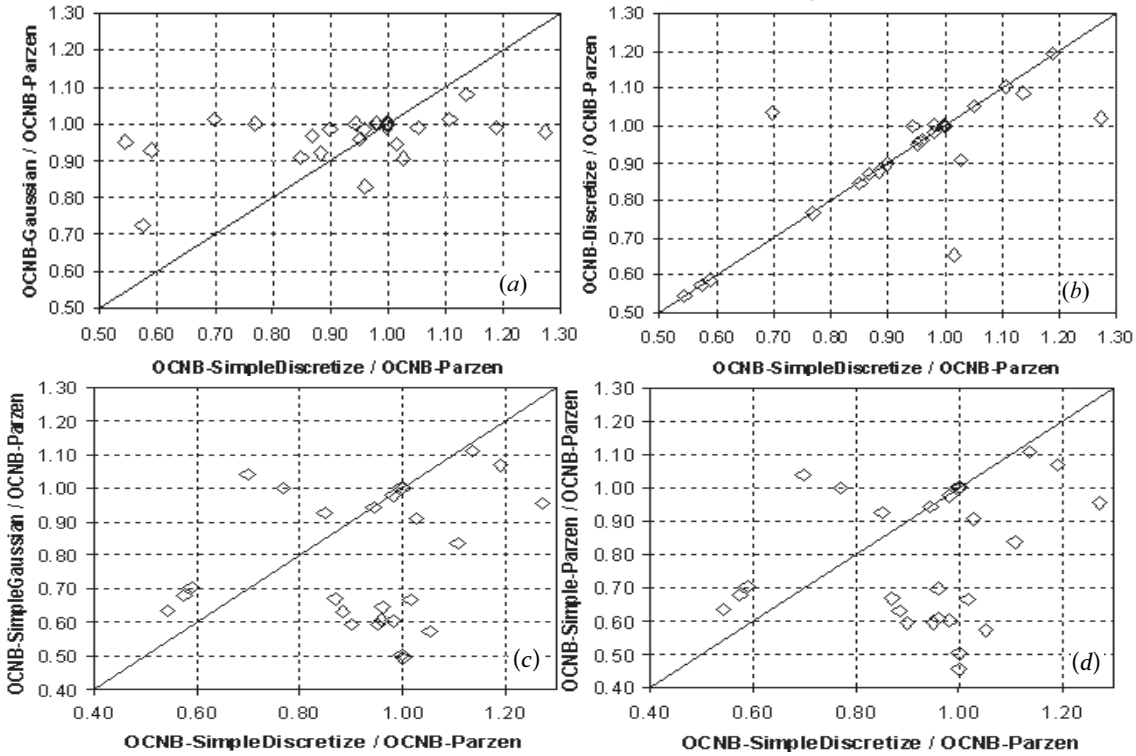

Figure 1. Relative perormance between OCNB-Parzen and other one-class Naïve Bayes classifiers.

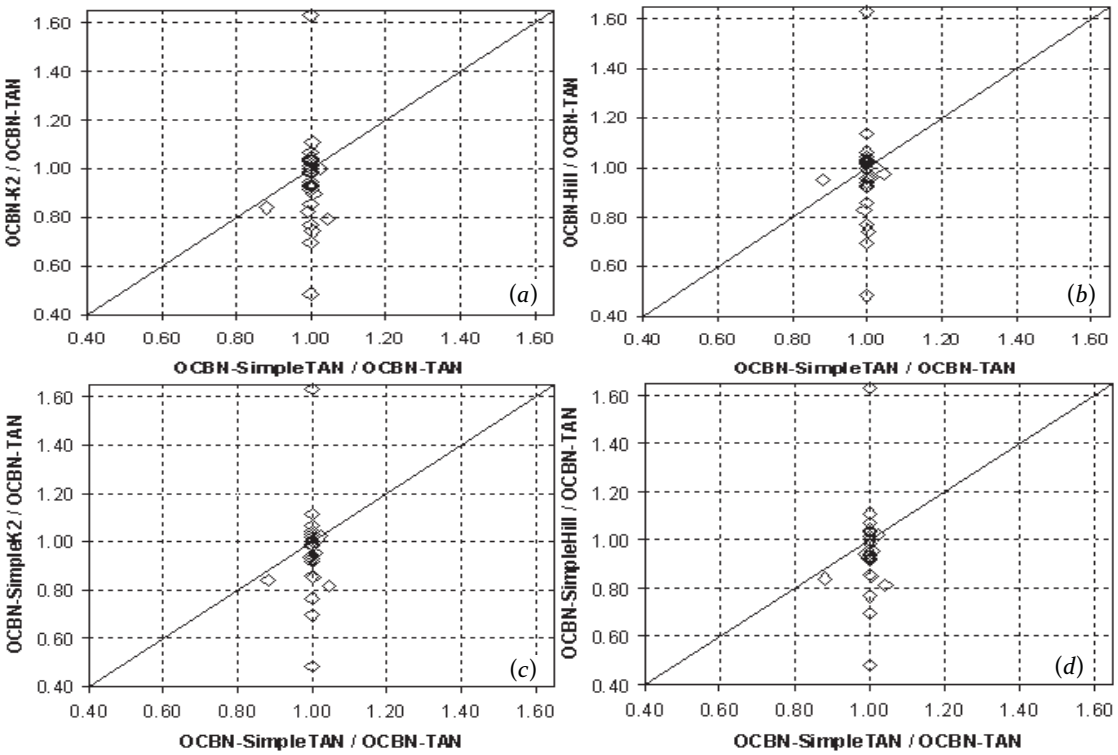

Figure 2. Relative performance between OCBN-TAN and other one-class Bayesian Network.

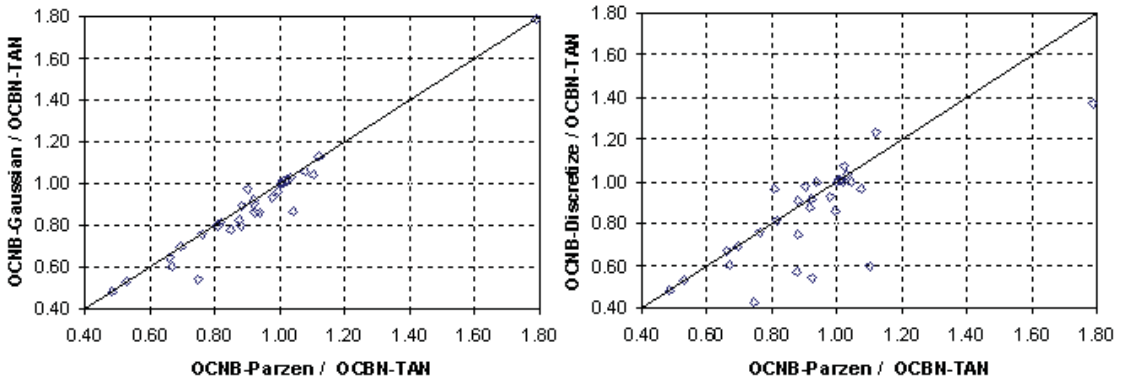

Figure 3. Relative peformance between improved OCBN-TAN and improved OCNBs (OCNB-Gaussian, OCNB-Parzen, and OCNB-Discretize).

Table 3. Summary of results for statistical tests between the improved OCNB-Parzen and all simple OCNB methods, and between the improved OCBN-TAN and all simple OCBN methods; the numbers in the string ".\.\" represent the chances of wins, ties, and loses of the improve one-class Bayes methods against the compared simple one-class Bayes methods.

	OCNB-SimpleGaussian	OCNB-SimpleDiscrete	OCNB-SimpleParzen	NB (Gaussian)
OCNB-Parzen	19\13\0	13\17\2	19\13\0	1\3\28
	OCBN-SimpleK2	OCBN-SimpleHill	OCNB-SimpleTAN	BN-K2
OCBN-TAN	16\12\4	18\10\4	1\30\1	1\5\26

Table 3. The same empirical result between the OCNB-Discrete and the OCNB-SimpleDiscrete can be found in (*b*) of Figure 2. The assumption in simple OCBNs might help one-class Bayes learning in nominal cases.

Our main observation is that the default TRR in the improved OCNB and the improved OCBN might not be proper in a noise circumstance. Tuning TRR can help learn an optimal one-class Bayes classifier. Instead of tuning TRR for training optimal OCNB and OCBN over the 30 benchmark datasets, we show experimental results on two real datasets: Ozone and OttawaRPB by tuning the TRR in Figure 4, where from (*a*) to (*c*) we drawn ROC curves for OCNB-Parzen, OCNB-SimpleDisretize (OCNB-S-D), OCBN-TAN, OCBN-SimpleTAN (OCBN-S-TAN), NB, and BN, which were built on Ozone.

In an ROC space, the point (0, 1) is denoted as the best performance while the diagonal line from the left bottom to the top right corners is denoted as a random classifier. The closer the curve is to the upper left corner, the better the classifier performs.

As we can see, when the TRR is set to the default 0.0 in (*a*) of Figure 4, two improved one-class Bayes classifiers: OCNB-Parzen and OCBN-TAN do not exhibit super performance while OCBN-S-D is worse than a random classifier. When TRR is set from 0.1 to 0.5, OCNB-Parzen is much optimized while OCBN-TAN unexpectedly degrades, and OCNB-S-D remains as the worst case and OCBN-S-TAN remains as a random classifier (no TRR). Further, experimental results on OttawaRPB by tuning optimal TRR is shown from (*d*) to (*f*) of Figure 4, where two improved one-class Bayes learning methods: OCNB-Parzen and OCBN-TAN are quite improved while two simple one-class Bayes learning methods: OCNB-S-D and OCBN-S-TAN perform as random classifiers (no TRR).

These observations show that the assumption of simple one-class Bayes learning has a restricted benefit for one-class learning. The improved method (e.g., OCBN-Parzen) can be better than the previously proposed simple one-class Bayes learning for probability estimation (e.g., OCBN-S-D) by tuning the TRR. However, from Figure 4, one-class Bayes classifiers such as OCNB-Parzen and OCBN-TAN are still inferior to the corresponding supervised learning methods, i.e., NB and BN, in two practical applications.

7. CONCLUSION AND FUTURE WORK

One-class Bayes learning consists of one-class Naïve Bayes and one-class Bayesian Network. It has been recognized that previously proposed one-class Bayes learning methods such as the simple one-class Naïve Bayes suffer from some limitations with the assumption that each nominal attribute value occurs at least one time in the underlying negative class distribution for probability estimation. We claim that it is ineffective on the domains with continuous attributes, and it is insufficient for probability estimation if the negative class distribution behaves complex, and the dependence on the negative class distribution is unreliable when the negative examples are unavailable. Further, the previous one-class Bayes method NBPC does not perform the probability estimation.

In this paper, we improve one-class Bayes learning by developing a new method for the probability calibration. The method learns the minimum probability score according to the target rejection rate, and the maximum probability score during the training time to help the probability estimation. The main advantages behind this new method are that it is independent of the negative class distribution and effective on various domains containing either nominal attributes or continuous attributes.

Our experimental results show that improved methods exhibit higher performance than simple methods on various domains containing nominal attribute and continuous attributes in most cases. In particular, in two practical applications, the improved one-class Bayes learning method is superior to simple one-class Bayes methods. This justifies the new probability calibration method for one-class Bayes learning.

When the improved one-class Bayes methods exhibit more successes than the previous one-class Bayes classifiers in practical applications, the main issue is that the current one-class Bayes learning methods cannot address a complex domain if there is a mixture probability model in the domain because they only build single classifier on the domain. Our study makes it possible to further improve one-class Bayes learning by assuming a possible Meta learning technique (like E2, an ensemble of positive example-based learning [26] or combining one-class classifiers [24]) such that one-class Bayes classifiers can be competitive with the traditional supervised learning methods for anomaly detection.

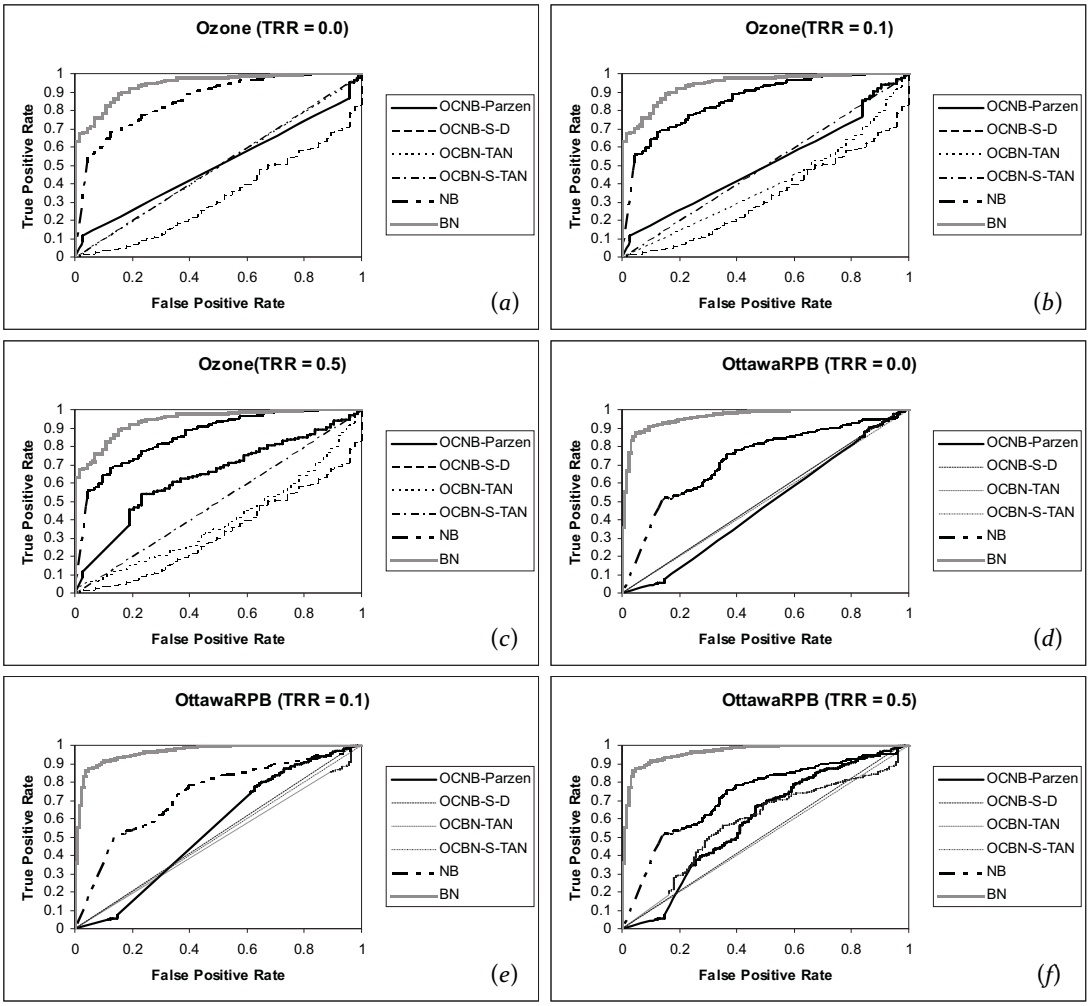

Figure 4. ROC curves of two improved Bayes classifiers: OCNB-Parzen and OCBN-TAN, and two simple one-class Bayes classifiers: OCNB-S-D and OCBN-S-TAN, and two supervised Bayes classifiers: NB and BN on Ozone and OttawaRPB. The figures from (*a*) to (*c*) are ROC curves of the classifiers built on Ozone with different TRRs; the figures from (*d*) to (*f*) are ROC curves of the classifiers built on OttawaRPB with different TRRs.

REFERENCES

[1] A. Asuncion and D. J. Newman. UCI Machine Learning Repository [http://www.ics.uci.edu/~mlearn/MLRepository.html]. Irvine, CA: University of California, School of Information and Computer Science. 2007.

[2] M. M. Breunig, H. P. Kriegel, R. T. NG, J. Sander. LOF: Identifying density-based local outliers. In Proceedings of the ACM SIGMOD International Conference on Management of Data, pp. 93--104. ACM Press (2000)

[3] V. Chandola, A. Banerjee, V. Kumar. Anomaly Detection: A Survey, vol. 41. ACM Computing Surveys (2009).

[4] V. Chandola, V. Mithal, V. Kumar. A Comparative Evaluation of Anomaly Detection Techniques for Sequence Data. TR 08-021 (2008)

[5] D. M. Chickering. Learning Bayesian networks is NP-complete. In D. Fisher & A. Lenz, *Learning from Data.* Springer-Verlag (1995).

[6] G. Cooper and E. Herskovits. A Bayesian method for the induction of probabilistic networks from data. Machine Learning, 9(4):309-347. (1992)

[7] C. Kruegel, D. Mutz, W. Robertson, and F. Valeur. Bayesian Event Classification for Intrusion Detection. Proceedings of the 19th Annual Computer Security Applications Conference. Page: 14, 2003.

[8] K. Das and J. Schneider. Detecting anomalous records in categorical datasets. In Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 220--229. ACM Press (2007)

[9] P. Datta: Characteristic Concept Representations. PhD thesis, University of California, Irvine (1997)

[10] R. O. Duda and P. E. Hart. Pattern Classification and Scene Analysis. A Wiley Intersience Publication (2000).

[11] C. Elkan. The foundations of cost-sensitive learning. Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence, pp. 973--978 (2001)

[12] T. Fawcett. *ROC graphs: Notes and practical considerations for data mining researchers.* Tech report HPL-2003-4. HP Laboratories, Palo Alto, CA, USA (2003)

[13] T. Fawcett and F. Provost. Adaptive Fraud Detection. Data Mining and Knowledge Discovery 1(3): 291--316 (1997)

[14] U. M. Fayyad and K. B. Irani. Multi-interval discretization of continuous-valued attributes for classification learning. In: Thirteenth International Joint Conference on Artificial Intelligence, 1022-1027, 1993.

[15] N. Friedman, D. Geiger, and M. Goldszmidt. Bayesian network classifiers. Machine Learning. 29(2-3):131-163 (1997).

[16] J. A. Hanley and B. J. McNeil. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 146:29--36 (1982)

[17] S. Hettich and S. D. Bay. The UCI KDD Archive [http://kdd.ics.uci.edu]. Irvine, CA: University of California, Department of Information and Computer Science (1999)

[18] K. Hempstalk and E. Frank. Discriminating Against New Classes: One-Class versus Multi-class Classification. Advances in Artificial Intelligence(AI2008), pp. 326--336 (2008)

[19] V. Hodge and J. Austin. A survey of outlier detection methodologies. Artif. Intel. Rev. 22(2) 85--126 (2004)

[20] D. Janakiram, V. Reddy, and A. Kumar. Outlier detection in wireless sensor networks using Bayesian belief networks. In Proceedings of the 1st International Conference on Communication System Software and Middleware, pp. 1--6 (2006)

[21] N. Japkowicz, C. Myers, and M. Gluck. A Novelty detection approach to classification. The proceedings of the 14th International conference on artificial Intelligence, pp. 518--523 (1995).

[22] B. Schölkopf, J. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson. Estimating the support of a high-dimensional distribution. Neural Comp., 13, 2001, pp. 1443--1471 (2001)

[23] D. M. J. Tax. One-class classification; concept-learning in the absence of counter-examples. Ph.D. thesis, Delft University of Technology (2001)

[24] D. M. J. Tax and R. P. W. Duin: Combining one-class classifiers. In: Kittler, J., Roli, F. (eds.) MCS 2001. LNCS, vol. 2096, pp. 299–308. Springer, Heidelberg (2001).

[25] K. Wang and S. J. Stolfo. One-class training for masquerade detection," 3rd IEEE Conference Data Mining and Workshop on Data Mining for Computer Security, pp. 1-10, 2003.

[26] C. P. Wei, H. C. Chen, and T. H. Cheng. Effective spam filtering: A single-class learning and ensemble approach. Decision Support Systems 45 (2008) 491–503.

[27] I. H. Witten, E. Frank. *Data Mining: Practical machine learning tools and techniques*, 2nd Edition, Morgan Kaufmann, San Francisco (2005)

[28] W. K. Wong, A. Moore, G. Cooper, and M. Wagner. Bayesian network anomaly pattern detection for disease outbreaks. In Proceedings of the 20th International Conference on Machine Learning, pp. 808--815. AAAI Press (2003)

[29] M. Yousef, N. Najami, and W. Khalifa. A comparison study between one-class and two-class machine learning for MicroRNA target detection. J. Biomedical Science and Engineering, 2010, 3, pp. 247--252 (2010)

[30] Zadrozny, B., Elkan, C.: Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers. Proceedings of the Eighteenth International conference on Machine Learning. Morgan Kaufmann (2001) 609-616.

[31] J. Zhang, H. Wang. Detecting outlying subspaces for high-dimensional data: The new task, algorithms, and performance. *Knowl. Inform. Syst. 10*, 3, pp. 333--355 (2006)

[32] J. Zhou, D. Dasgupta. V-detector: An efficient negative selection algorithm with ‘‘probably adequate” detector coverage. Information Sciences 179 (2009) 1390--1406. (2009)

[33] K. Zhang, W. Fan, X. J. Yuan, I. Davidson, and X. S. Li. Forecasting Skewed Biased Stochastic Ozone Days: Analyses and Solutions. In Proceedings of the International Conference on Data Ming, pp. 753-764. 2006.