Umezu *et al. Pacific Journal of Mathematics for Industry* (2015) 7:8
DOI 10.1186/s40736-015-0019-z

Pacific Journal of Mathematics for Industry
a SpringerOpen Journal

ORIGINAL ARTICLE Open Access

Defect rate evaluation via simple active learning

Yuta Umezu[1], Hidetoshi Matsuoka[2], Hiroshi Ikeda[2] and Yoshiyuki Ninomiya[3*]

Abstract

In the preparatory stage of product manufacturing, its defect risk is often evaluated by checking whether experimentally manufactured products cause the defect or not. The experimentally manufacturing is conducted for various values of variables which may related the defect, but manufacturing products for all combinations of the values will cost a lot especially when the number of variables is large. To overcome this problem, active learning methods which may be able to evaluate the defect risk efficiently by selecting values purposefully are considered. In this paper, it is pointed out that even a modern active learning method is inappropriate if the nonlinearity of the relation between the variables and the defect is strong and if the defect rate is small. And then a simple active learning method which can work well for such a case is proposed. Through simulation studies and real data analysis, the validity of the proposed method is checked.

Keywords: Importance sampling; Nonlinear classification; Optimal design; Random sampling; Rare event; Support vector machine

1 Introduction

Let us consider a product with the risk of having a defect at the time of manufacturing. We assume that the risk depends on the values of various variables such as the temperature, the amount of an ingredient or operating time. In particular, letting $y_i = -1$ and $y_i = 1$ mean that the i-th product respectively has and does not have a defect and letting $\boldsymbol{x}_i$ ($\in \mathcal{D} \subset \mathbb{R}^p$) be the values of such p variables, we assume that the defect risk for $\boldsymbol{x}_i$ is given by

$$\mathrm{P}(y_i = -1|\boldsymbol{x}_i) = \frac{\exp(h(\boldsymbol{x}_i))}{1 + \exp(h(\boldsymbol{x}_i))}$$

for an appropriate function $h(\boldsymbol{x})$. In product manufacturing, it is indispensable to evaluate $\mathrm{P}(y_i = -1|\boldsymbol{x}_i)$ because we can avoid to yield defects which may cause a severe damage in manufacturing company if we know $\mathrm{P}(y_i = -1|\boldsymbol{x}_i)$ (see e.g., Katayama et al. [3], Sun & Li [11]). Let us consider the appropriate function as $h(\boldsymbol{x})$. Needless to say, the products are manufactured not to have any defect, and so y tends to be one if $\boldsymbol{x}$ is in the central zone of the domain $\mathcal{D}$. That is, the region $\{\boldsymbol{x} \mid h(\boldsymbol{x}) > 0\}$ should be small and at the edge of $\mathcal{D}$. In addition, considering that the cause of the defect is usually multiple, such regions, which produces a defect more likely than not, tend to be scattered at the edge of $\mathcal{D}$. Namely, such regions should not only have a strongly nonlinear boundary but also tend to be separated although they will not be far from each other. Therefore, we suppose that $h(\boldsymbol{x})$ has a strong nonlinearity although it will not have a drastic fluctuation.

Under this situation, we consider to estimate the decision rule $\mathrm{sgn}(h(\boldsymbol{x}))$ along with the defect risk by sampling $\{(y_i, \boldsymbol{x}_i) \mid \boldsymbol{x}_i \in \mathcal{X};\ i = 1, 2, \ldots, n\}$ appropriately. Here $\mathcal{X}$ ($\subset \mathcal{D}$) is a set of candidates of $\boldsymbol{x}_i$ we can sample. This is $\mathcal{D}$ itself in some cases and a finite set in other cases. As a versatile method for giving nonlinear decision rules, recently the SVM (support vector machine; e.g., Cristianini & Shawe-Taylor [2], Scholköpf & Smola [9]) becomes a standard tool. In addition, the Gaussian process regression method (e.g., Rasmussen & Williams [8]) is known to have comparable performance. Although these methods are capable of dealing with strong nonlinearity, a lot of samples are required to deal with it as a matter of course. This requirement becomes evident when the dimension of $\boldsymbol{x}$ is large. Therefore, an appropriate selection of samples from $\mathcal{X}$ is important to estimate $\mathrm{sgn}(h(\boldsymbol{x}))$ efficiently for the case where the sampling cost is not necessarily low and

*Correspondence: nino@imi.kyushu-u.ac.jp
[3]Institute of Mathematics for Industry, 744, Motooka, Nishi-ku, Fukuoka 819-0395, Japan
Full list of author information is available at the end of the article

we can select samples purposefully. This type of appropriate selection is called optimal design in classic statistical area and active learning in machine learning area.

While the active learning versions for the above-mentioned nonlinear discriminant methods are not sufficiently developed, the ASVM (active learning SVM) proposed by Tong & Koller [12] gets a lot of attention. The ASVM, an active learning method specialized for the SVM, can be implemented easily for considerably huge data and has a high computational efficiency. Roughly speaking, however, the sampling scheme in the ASVM is to select samples close to the decision boundary estimated by already gotten data, and it is often the case that the ASVM does not work well for our problem owing to the sampling scheme. Also the method in Umezu & Ninomiya [13] does not work well for our problem although it was proposed to overcome the weak point of the ASVM. Our main purpose is to show the lack of the development in active learning methods for a kind of discriminant problems even though they are simple and general problems. In addition, as the first step of the development, we try to provide a simple and computer-efficient method capable of such discriminant problems.

The rest of the paper is as follows. In Sections 2 and 3, we will introduce the above-mentioned non-linear discriminant methods and active learning methods, respectively. In Section 4, after explaining why such methods are not suitable for our problem, we will propose a simple but effective method. The method is shown to be valid through a simulation study in Section 5, and then we report the result in which the method is applied to real data in Section 6. We suggest how to evaluate the rate of producing defects in Section 7, and some concluding remarks are presented in Section 8.

2 Existing discriminant methods

Nowadays the SVM is one of the most standard tools as an efficient and effective non-linear discriminant method. Let us explain its detail because later we not only introduce its active learning version but also use it in our proposal.

The SVM is a classifier whose decision rule is

$$y = \mathrm{sgn}(\boldsymbol{w} \cdot \boldsymbol{\phi}(\boldsymbol{x}))$$

for unknown input $\boldsymbol{x}$, where $\mathrm{sgn}(\cdot)$ is the sign function that returns 1 or -1 when the argument is respectively positive or negative, and $\boldsymbol{w} \cdot \boldsymbol{\phi}(\boldsymbol{x})$ is called a discriminant function. Here, $\boldsymbol{\phi}(\cdot)$ is a map from the space $\mathcal{X}$ of the input $\boldsymbol{x}$ to a higher dimensional so-called feature space $\mathcal{F}$, which satisfies $\boldsymbol{\phi}(\boldsymbol{x}) \cdot \boldsymbol{\phi}(\tilde{\boldsymbol{x}}) = k(\boldsymbol{x}, \tilde{y}x)$, where $k(\cdot,\cdot)$ is a symmetric and positive definite kernel. In addition, $\boldsymbol{w}$ ($\in \mathcal{F}$) is an unknown coefficient for $\boldsymbol{\phi}(\boldsymbol{x})$.

The optimal coefficient $\hat{\boldsymbol{w}}$ is given by maximizing a so-called margin. For a dataset of n-tuple $\{(y_i, \boldsymbol{x}_i) \mid i = 1, 2, \ldots, n\}$ consisting of an input $\boldsymbol{x}_i$ ($\in \mathcal{D}$) and its output y_i ($\in \{1, -1\}$), the maximization problem reduces to

$$\max_{\boldsymbol{w} \in \mathcal{V}} \min_{i \in \{1,2,\ldots,n\}} \{y_i \boldsymbol{w} \cdot \boldsymbol{\phi}(\boldsymbol{x}_i)\},$$

where

$$\mathcal{V} \equiv \{\boldsymbol{w} \in \mathcal{F} \mid ||\boldsymbol{w}|| = 1;\ \forall i,\ y_i \boldsymbol{w} \cdot \boldsymbol{\phi}(\boldsymbol{x}_i) > 0\} \tag{1}$$

is called the version space. You may think this optimization problem is hard to solve because $\mathcal{F}$ is a high dimensional space, but it can be shown by the representer theorem (e.g., Shawe-Taylor & Cristianini [10]) that the optimal coefficient of $\boldsymbol{w}$ provides a simple estimated discriminant function as

$$\hat{\boldsymbol{w}} \cdot \boldsymbol{\phi}(\boldsymbol{x}) = \sum_{i=1}^{n} \hat{\alpha}_i k(\boldsymbol{x}, \boldsymbol{x}_i), \tag{2}$$

where $\hat{\alpha}_i$ is given by the following optimization problem:

$$\max_{\boldsymbol{\alpha} \in \mathbb{R}^n} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j}^{n} \alpha_i \alpha_j y_i y_j k(\boldsymbol{x}_i, \boldsymbol{x}_j)$$

$$\text{subject to} \quad \sum_{i=1}^{n} \alpha_i y_i = 0 \ \text{ and } \ \forall i,\ \alpha_i \geq 0. \tag{3}$$

Since this optimization problem is convex with respect to the variables to be optimized, we can use a popular method of convex optimization (e.g., Boyd & Vandenberghe [1]). Note that in general $\boldsymbol{\phi}(\boldsymbol{x})$ is nonlinear with respect to $\boldsymbol{x}$, the discriminant function $\hat{\boldsymbol{w}} \cdot \boldsymbol{\phi}(\boldsymbol{x})$ and the decision boundary $\{\tilde{\boldsymbol{x}} \mid \hat{\boldsymbol{w}} \cdot \boldsymbol{\phi}(\tilde{\boldsymbol{x}}) = 0\}$ are also nonlinear.

While there are so many kinds of kernels (e.g., Chapter 4 in Rasmussen & Williams [8]), we can say that one of the most frequently used kernels is Gaussian kernel with the form of

$$k(\boldsymbol{x}, \tilde{\boldsymbol{x}}) = \exp(-\gamma \|\boldsymbol{x} - \tilde{\boldsymbol{x}}\|^2) \ \ (\boldsymbol{x}, \tilde{\boldsymbol{x}} \in \mathbb{R}^p),$$

where γ (> 0) is a tuning parameter which controls the dispersion of the kernel. In this paper, we use this kernel and select the optimal value of γ by cross-validation.

Except for the SVM, the Gaussian process regression method adapted for discriminant problems is well-known as a nonlinear discriminant method which has a similar performance to the SVM with the Gaussian kernel (e.g., Rasmussen & Williams [8]). In this method, a monotone function of the probability of being $y = 1$ ($y = -1$) for the input $\boldsymbol{x}$, which is denoted by $Z(\boldsymbol{x})$, is regarded as the Gaussian process (random field) with a positive autocorrelation. If the autocorrelation $\mathrm{Cor}[Z(\boldsymbol{x}), Z(\tilde{\boldsymbol{x}})]$ and the relationship between $\mathrm{P}[y = 1|\boldsymbol{x}]$ and $Z(\boldsymbol{x})$ is given, about the output $\tilde{y}$ for an unknown input $\tilde{\boldsymbol{x}}$, we can evaluate the distribution of $\tilde{y}$ conditional on $\{(\boldsymbol{x}_i, y_i) \mid i = 1, 2, \ldots, n\}$ and $\tilde{\boldsymbol{x}}$ in a simple form. Then, we can predict $\tilde{y}$ using

the conditional distribution. Because this prediction is based on a framework of classic statistics, we can evaluate a prediction accuracy, conduct a variable selection and implement an optimal design according to the framework.

3 Active learning

The active learning method (optimal design method) is to design inputs to improve a learning (estimation) accuracy for the case where we can design the inputs purposefully. For the SVM, classic optimal design methods are not applicable, and then a specialized method is developed. In this section, we introduce such a method, the ASVM, proposed by Tong & Koller [12].

Before introducing the ASVM, we first describe classic optimal design methods, the A-optimal and D-optimal designs. Generally speaking, for a parameter vector $\boldsymbol{\theta}$ and its estimator $\hat{\boldsymbol{\theta}}$, the mean squared error matrix $\mathrm{E}[\,(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})'\,]$ is a natural index for the estimation accuracy. It is divided into the variance-related term $\mathrm{E}[\,(\hat{\boldsymbol{\theta}}-\mathrm{E}[\,\hat{\boldsymbol{\theta}}\,])(\hat{\boldsymbol{\theta}}-\mathrm{E}[\,\hat{\boldsymbol{\theta}}\,])'\,]$ and the bias-related term $(\mathrm{E}[\,\hat{\boldsymbol{\theta}}\,]-\boldsymbol{\theta})(\mathrm{E}[\,\hat{\boldsymbol{\theta}}\,]-\boldsymbol{\theta})'$, and in well-used estimation methods such as the maximum likelihood method, the former becomes the main term asymptotically. In classic optimal design methods, it is proposed to give a new input which minimize the trace or determinant of the main term $\mathrm{E}[\,(\hat{\boldsymbol{\theta}}-\mathrm{E}[\,\hat{\boldsymbol{\theta}}\,])(\hat{\boldsymbol{\theta}}-\mathrm{E}[\,\hat{\boldsymbol{\theta}}\,])'\,]$, and it is called A-optimal design or D-optimal design, respectively (Kiefer [4, 5], Kiefer & Wolfowitz [6]). Because we cannot evaluate this matrix explicitly in general, it is common to use the inverse of Fisher's information matrix in place of it, which is asymptotically equivalent to it under some regularity conditions. Note that the A-optimal and D-optimal designs are equivalent under a setting of linear regression. Also note that in the D-optimal design, the input giving the maximum prediction variance of its output is selected under some conditions, which is an important property. That is, letting $\hat{y}(\boldsymbol{x};\hat{\boldsymbol{\theta}})$ be the predictive value of the output for $\boldsymbol{x}$, the D-optimal design selects $\mathrm{argmax}_{\boldsymbol{x}}\mathrm{V}[\,\hat{y}(\boldsymbol{x};\hat{\boldsymbol{\theta}})]$ as a new input, and so it is regarded as a method which gradually reduces region which gives unstable prediction.

On the other hand, the SVM cannot use the above optimal design methods because we have no evaluation formula for the variances of parameter estimators in the SVM setting. Actually as seen from (3), the numbers of parameters and samples are the same, and so we have no evaluation formula even in an asymptotical form. Under this situation, Tong & Koller [12] propose a new criterion for sampling based on the version space for the SVM.

After getting a dataset of n-tuple $\{(y_i,\boldsymbol{x}_i)\mid i=1,2,\ldots,n\}$, the version space is defined as in (1). In this definition, each $y_i\boldsymbol{w}\cdot\boldsymbol{\phi}(\boldsymbol{x}_i)>0$ represents a half space in $\mathcal{F}$ and $\mathcal{V}$ represents the polyhedral body which is the product set of the half spaces. As the $(n+1)$-th new sample, they propose to select $\boldsymbol{x}_{n+1}$ $(\in\mathcal{X})$ such that the hyperplane $\boldsymbol{w}\cdot\boldsymbol{\phi}(\boldsymbol{x}_{n+1})=0$ divides $\mathcal{V}$ into two parts as equally as possible. It is indicated in Tong & Koller [12] that the $(n+1)$-th new sample is close to

$$\operatorname*{argmin}_{\boldsymbol{x}\in\mathcal{X}}|\hat{\boldsymbol{w}}\cdot\boldsymbol{\phi}(\boldsymbol{x})|$$

for the estimated discriminant function in (2). Therefore, this method selects $\boldsymbol{x}_{n+1}$ such that $\hat{\boldsymbol{w}}\cdot\boldsymbol{\phi}(\boldsymbol{x}_{n+1})$ is close to 0, in other words, $\boldsymbol{x}_{n+1}$ which is close to the decision boundary.

4 Proposed method

In this section, first we will point out a severe problem in the ASVM under our situation. Next we will propose a simple active learning method which avoids the problem.

For simplicity, we assume that the dimension p of inputs is 2, and let us consider the example in which all candidates of the inputs and their outputs are as in the left of Fig. 1. We consider the situation where the region of $\boldsymbol{x}$ in which $y=-1$ is given more likely than $y=1$ tends to be separated and near the edge but they will not be far from each other. As written in Section 1, this situation is natural for the defect rate evaluation problem in product manufacturing. Actually, our real data treated later have this situation while p is much larger.

Let us imagine that the ASVM is applied to this example. Although the ASVM is an active learning method, we cannot get samples actively in the first stage, and so we get them completely at random. Suppose that they are filled points in the right of Fig. 1. If we estimate the decision boundary based on them by the SVM, the curve in the figure is obtained. After that, we get samples according to the sampling scheme of the ASVM. Then decision boundary for the left group of $y=-1$ will be improved step by step because inputs close to the estimated decision boundary must be selected as explained in Section 3. On the other hand, the inputs with $y=-1$ in the right group are rarely sampled, and so the estimated decision boundary near the right group will not appear for a long time. Thus, on the whole, the discriminant accuracy will not be much improved even if the sampling is repeated.

When p is large and the number of outputs of $y=-1$ is small, the above phenomenon becomes apparent. The region giving $y=-1$ more likely than not tends to be more separated, and it is difficult to get any sample from a number of separated regions at the first random sampling. In addition, it takes longer time to get a sample from the separated regions where no inputs are sampled at the first stage. Thus, it is indicated that active learning methods which get samples near the estimated decision boundary are not suitable for this type of cases.

On the other hand, the Gaussian process regression method, which has comparable performance to the SVM,

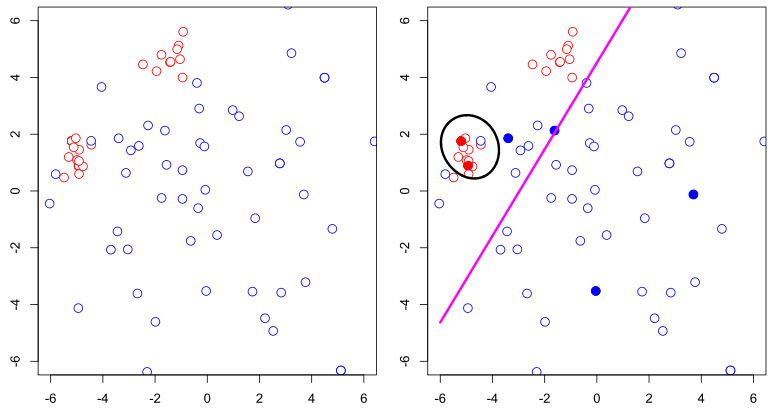

Fig. 1 Example of two-dimensional case. In the left, the blue and red open circles represent the candidates of inputs with $y = 1$ and $y = -1$, respectively. In the right, the filled circles and the curve represent the sampled inputs, the black solid line is the estimated decision boundary via SVM, and the purple solid line is the linear discriminant boundary

is based on a framework of classic statistics, and so the prediction accuracy can be evaluated. Considering the important property of the D-optimal design explained in Section 3, Umezu & Ninomiya [13] proposed a new optimal design method which selects samples with the maximum prediction instability measured by an entropy. Using this method, we can select inputs by considering its closeness to not only the estimated decision boundary but also already sampled inputs. However, this can be regarded as a method between the ASVM and the method with sampling completely at random, and so it must be inappropriate for our problem.

Hence, we consider a method which does not depend on the estimated decision boundary. After sampling an input uniformly at random from $\mathcal{X}$ and obtaining an output according to the input, which is repeated till at least one output with $y = -1$ is obtained, we consciously forget the nonlinearity of our discriminant problem and conduct a linear discriminant analysis. Let $\tilde{h}(\boldsymbol{x})$ be the linear discriminant function, and let $\mathcal{D}^- = \{\boldsymbol{x} \mid \tilde{h}(\boldsymbol{x}) < 0\}$. In this paper, we consider a hyperplane which consists of the points such that the distances from the centers of $\mathcal{D}$ and of inputs with $y = -1$ are equal, and we define $\tilde{h}(\boldsymbol{x})$ so that $\{\boldsymbol{x} \mid \tilde{h}(\boldsymbol{x}) = 0\}$ is the hyperplane. Because the separated regions giving $y = -1$ more likely than not are not large, not far from each other, and at the edge, it can be expected that most of such regions are included in $\mathcal{D}^-$ (see the right in Fig. 1). Then we sample inputs uniformly at random from $\tilde{\mathcal{X}} \equiv \mathcal{D}^- \cap \mathcal{X}$ and obtain outputs according to the inputs. By repeating this procedure, we can expect to get samples from all the separated regions. In this situation, since the area of $\mathcal{D}^-$ is not large in comparison with the area of $\mathcal{D}$, we will be able to get inputs with $y = -1$ efficiently. Finally we recall the nonlinearity of our discriminant problem, and then estimate the discriminant function $\hat{\boldsymbol{w}} \cdot \boldsymbol{\phi}(\boldsymbol{x})$ by applying the SVM. This procedure can be summarized as in Table 1.

5 Simulation study

To compare "Linear discrimination"-based active learning with the SVM (LSVM) proposed in Section 4, the method in Tong & Koller [12] (ASVM) and "Random sampling" with the SVM (RSVM), we conduct a simulation study in this section. In the RSVM, we sample inputs from $\mathcal{X}$ completely at random without active learning and finally use the SVM to estimate the discriminant function. Because we must apply these methods many times in the simulation study, we set the dimension of inputs and the number of sampled inputs are small.

Concretely speaking, first we produce 2,000 inputs with a negative output by

$$\boldsymbol{x} \sim \mathrm{Mix}(1/2, \mathrm{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1), \mathrm{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)) \;\Rightarrow\; y = -1$$

and 98,000 inputs with a positive output by

$$\boldsymbol{x} \sim \mathrm{N}(\boldsymbol{\mu}_3, \boldsymbol{\Sigma}_3) \;\Rightarrow\; y = 1,$$

Table 1 Procedure in LSVM

1)	For $i = 1, 2, \ldots, N$, sample $\boldsymbol{x}_i$ uniformly at random from $\mathcal{X}$, and obtain $y_i \in \{\pm 1\}$ according to $\boldsymbol{x}_i$.
2)	Obtain a linear discriminant function $\tilde{h}(\boldsymbol{x})$ such that $\{\boldsymbol{x} \mid \tilde{h}(\boldsymbol{x}) = \boldsymbol{0}\}$ becomes the hyperplane which is equidistant from the centers of $\mathcal{D}$ and of inputs with $y = -1$.
3)	Set $\tilde{\mathcal{X}} = \{\boldsymbol{x} \in \mathcal{X} \mid \tilde{h}(\boldsymbol{x}) < 0\}$
4)	For $i = 1, 2, \ldots, M$, sample $\tilde{\boldsymbol{x}}_i$ uniformly at random from $\tilde{\mathcal{X}}$, and obtain $\tilde{y}_i \in \{\pm 1\}$ according to $\tilde{\boldsymbol{x}}_i$
5)	Iterate 2) to 4) for K times
6)	Estimate a discriminant function $\hat{\boldsymbol{w}} \cdot \boldsymbol{\phi}(\boldsymbol{x})$ by SVM

and pool them. Here, $\mathrm{Mix}(1/2, \mathrm{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1), \mathrm{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2))$ means the mixture distribution of $\mathrm{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $\mathrm{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ with the mixing rate $1 : 1$. Letting $\boldsymbol{R}(\theta)$ be the two dimensional rotation matrix with the angle θ, we set

$$\boldsymbol{\mu}_1 = \begin{pmatrix} 0 \\ 5 \end{pmatrix},\ \boldsymbol{\Sigma}_1 = \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix}$$

$$\boldsymbol{\mu}_2 = \boldsymbol{R}(\theta)\boldsymbol{\mu}_1,\ \boldsymbol{\Sigma}_2 = \boldsymbol{R}(\theta)\boldsymbol{\Sigma}_1\boldsymbol{R}(\theta)'$$

$$\boldsymbol{\mu}_3 = \begin{pmatrix} 0 \\ 0 \end{pmatrix},\ \boldsymbol{\Sigma}_3 = \begin{pmatrix} 5 & 0 \\ 0 & 5 \end{pmatrix}$$

as the values of the parameters. The inputs with $y = -1$ form two groups and the angle θ indicates their distance. Next we compare the methods by getting samples from the pooled data. In every method, first we get 50 samples completely at random, and then we iterate 25 times of samplings in which we get 10 samples at one time according to the procedure of each method. That is, in Table 1 for the LSVM, we set $N = 50$, $M = 10$ and $K = 25$.

In Table 2, by each designed value of $\left(\theta, \sigma_1^2, \sigma_2^2\right)$, we can check the transitions of FPRs (false positive rates) caused by increasing the number of iterations of sampling, where the FPR is defined by $\#\{\, i \mid \hat{\boldsymbol{w}} \cdot \boldsymbol{\phi}\,(\boldsymbol{x}_i) > 0,\ y_i = -1\}/\#\{i \mid y_i = -1\}$. Here we do not report about the FNRs (false negative rates) because the FPR is more important to be checked than the FNR in our problem and because the FNRs for all methods were always very close to one and almost the same values. The values in the table are the averages and standard deviations of FPRs based on 50 simulations for each method. In every method, the FPR is decreasing when the number of iterations is increased.

First, it can be seen in every case that the RSVM provides much higher values of FPRs than those of the LSVM and the ASVM. This is because the RSVM can rarely get samples with $y = -1$ unlike the other two methods. Next, it can be seen by comparing the two methods that basically the LSVM is superior to the ASVM when the number of iterations become large while the ASVM is superior to the LSVM when it is small. This is because the ASVM can quickly get samples with $y = -1$ close to the initially gotten sample with $y = -1$ but cannot get those far from it. For the case where the two groups of the inputs with $y = -1$ are close, e.g., $\theta = \pi/9$, the ASVM has a possibility of finding any of those, and so the two methods are comparable. In addition, for the case where the two groups are too far from each other, e.g., $\theta = 4\pi/9$, even the LSVM does not have a possibility of finding any of inputs with $y = -1$, and so the superiority of the LSVM becomes small. For the other cases, the LSVM is clearly better than the ASVM.

Table 2 Transition of FPRs for simulated data

		Number of iterations				
θ	Method	5	10	15	20	25
(a) Case of $\left(\sigma_1^2, \sigma_2^2\right) = (0.1, 0.3)$						
$\pi/9$	LSVM	0.912	0.022	0.016	0.021	0.031
	sd	0.188	0.020	0.007	0.011	0.014
	ASVM	0.102	0.027	0.036	0.037	0.026
	sd	0.132	0.017	0.019	0.016	0.009
	RSVM	1.000	1.000	0.932	0.654	0.352
	sd	0.000	0.001	0.169	0.332	0.299
$2\pi/9$	LSVM	0.994	0.360	0.068	0.043	0.045
	sd	0.024	0.278	0.096	0.026	0.026
	ASVM	0.540	0.503	0.503	0.504	0.504
	sd	0.080	0.002	0.003	0.006	0.002
	RSVM	1.000	1.000	0.988	0.943	0.805
	sd	0.000	0.000	0.061	0.127	0.227
$\pi/3$	LSVM	0.979	0.439	0.085	0.033	0.027
	sd	0.084	0.244	0.154	0.098	0.098
	ASVM	0.532	0.502	0.503	0.503	0.505
	sd	0.063	0.002	0.004	0.002	0.004
	RSVM	1.000	1.000	0.998	0.969	0.833
	sd	0.000	0.000	0.013	0.123	0.255
$4\pi/9$	LSVM	0.996	0.517	0.337	0.281	0.269
	sd	0.028	0.177	0.227	0.250	0.247
	ASVM	0.528	0.503	0.503	0.504	0.504
	sd	0.037	0.003	0.003	0.003	0.003
	RSVM	1.000	1.000	1.000	0.940	0.841
	sd	0.000	0.000	0.000	0.178	0.239
(b) Case of $\left(\sigma_1^2, \sigma_2^2\right) = (0.2, 0.2)$						
$\pi/9$	LSVM	0.787	0.032	0.027	0.028	0.029
	sd	0.349	0.024	0.018	0.015	0.017
	ASVM	0.083	0.028	0.031	0.026	0.018
	sd	0.107	0.016	0.020	0.016	0.009
	RSVM	0.900	0.900	0.850	0.663	0.286
	sd	0.303	0.303	0.322	0.362	0.257
$2\pi/9$	LSVM	1.000	0.359	0.080	0.047	0.050
	sd	0.003	0.280	0.109	0.038	0.041
	ASVM	0.528	0.503	0.503	0.504	0.505
	sd	0.043	0.002	0.004	0.003	0.003
	RSVM	1.000	1.000	0.994	0.941	0.798
	sd	0.000	0.000	0.037	0.129	0.262
$\pi/3$	LSVM	1.000	0.506	0.080	0.028	0.019
	sd	0.000	0.291	0.162	0.098	0.070
	ASVM	0.524	0.502	0.502	0.503	0.503
	sd	0.037	0.001	0.002	0.003	0.003
	RSVM	1.000	1.000	0.992	0.946	0.791
	sd	0.000	0.000	0.049	0.133	0.261

Table 2 Transition of FPRs for simulated data *(Continued)*

$4\pi/9$	LSVM	0.990	0.512	0.239	0.192	0.182
	sd	0.052	0.180	0.228	0.244	0.241
	ASVM	0.535	0.503	0.502	0.503	0.504
	sd	0.053	0.003	0.002	0.002	0.002
	RSVM	1.000	1.000	1.000	0.978	0.821
	sd	0.000	0.000	0.000	0.072	0.232
(c) Case of $(\sigma_1^2, \sigma_2^2) = (0.3, 0.1)$						
$\pi/9$	LSVM	0.901	0.046	0.033	0.030	0.024
	sd	0.221	0.034	0.015	0.014	0.011
	ASVM	0.061	0.029	0.022	0.016	0.011
	sd	0.060	0.018	0.011	0.008	0.005
	RSVM	1.000	1.000	0.941	0.649	0.401
	sd	0.000	0.000	0.159	0.338	0.347
$2\pi/9$	LSVM	0.970	0.333	0.055	0.042	0.065
	sd	0.096	0.236	0.061	0.017	0.029
	ASVM	0.509	0.476	0.475	0.481	0.482
	sd	0.046	0.111	0.110	0.091	0.092
	RSVM	1.000	1.000	0.995	0.943	0.834
	sd	0.000	0.000	0.032	0.138	0.255
$\pi/3$	LSVM	0.991	0.471	0.082	0.030	0.034
	sd	0.061	0.245	0.137	0.068	0.068
	ASVM	0.513	0.503	0.502	0.504	0.503
	sd	0.017	0.002	0.001	0.003	0.004
	RSVM	1.000	1.000	1.000	0.944	0.827
	sd	0.000	0.000	0.000	0.157	0.253
$4\pi/9$	LSVM	0.991	0.493	0.241	0.179	0.159
	sd	0.042	0.200	0.237	0.235	0.227
	ASVM	0.531	0.504	0.503	0.503	0.503
	sd	0.075	0.002	0.002	0.002	0.002
	RSVM	1.000	1.000	0.991	0.971	0.852
	sd	0.000	0.000	0.061	0.101	0.220

6 Real data analysis

In this section, we compare the methods through applying them to some trial data which is used in a real product manufacturing. The data consists of 97,740 samples with $y = 1$ and 2,260 samples with $y = -1$, and the dimension p of the inputs is 18. As in the situation we treated until now, the inputs with $y = -1$ form several groups at the edge of the domain $\mathcal{D}$. Note that we know all values of the outputs because this data is for trial. Using these known values, we can estimate a good discriminant function without active learning, but here we suppose to know only the values of outputs gotten by sampling. Needless to say, it is because we look ahead to apply the methods to non-trial data.

In every method, first we get 500 samples completely at random, and then we iterate 200 times of samplings in which we get 50 samples at one time according to the procedure of each method. That is, in Table 1 for the LSVM, we set $N = 500$, $M = 50$ and $K = 200$. In Fig. 2, we plot the transition of the FPR for each method, which is measured by making test data from non-sampled data with $y = -1$. It can be seen that the values of the FPR for the LSVM are always smaller than those of the ASVM and become stable after about 50 times iterations while those of the ASVM are decreasing slowly. About the RSVM, the values of the FPR become temporally smaller than those of the LSVM, but it will be by accident because the values considerably fluctuate after that. Moreover, the RSVM is superior to the ASVM in this case. It may be because that there are too many groups of the inputs with $y = -1$ to deal with by the ASVM. Actually the values do not become stable even after 100 times iterations.

7 Evaluation of defect rate

While efficient estimation of the discriminant function for a defect was discussed until now, it is often the case actually in product manufacturing that the estimation of its defect rate has more concern. Then we consider to evaluate

$$\mathrm{E}[y = -1] = \int \rho(\boldsymbol{x}) f(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}, \tag{4}$$

where $f(\boldsymbol{x})$ is the probability density function of $\boldsymbol{x}$, and $\rho(\boldsymbol{x})$ is the probability of being $y = -1$ at $\boldsymbol{x}$.

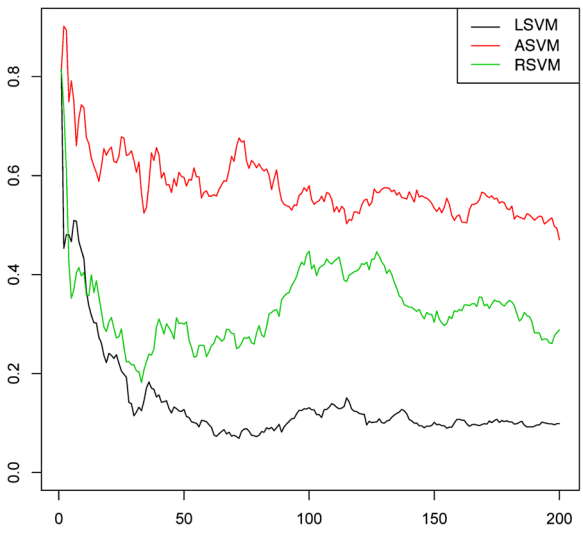

Fig. 2 Transition of FPRs for real data. The horizontal axis indicates the number of sampling iterations and the vertical axis indicates the FPR

First we model the local defect rate by

$$\rho(\boldsymbol{x}) = \frac{\exp(a\hat{\boldsymbol{w}} \cdot \boldsymbol{\phi}(\boldsymbol{x}) + b)}{1 + \exp(a\hat{\boldsymbol{w}} \cdot \boldsymbol{\phi}(\boldsymbol{x}) + b)} \tag{5}$$

using the discriminant function $\hat{\boldsymbol{w}} \cdot \boldsymbol{\phi}(\boldsymbol{x})$ obtained by the SVM (e.g., Platt [7]). Here a and b are unknown parameters, and we estimate them by the maximum likelihood method under the setting where y_i is an independent sample from the Bernoulli distribution $\mathrm{Be}(\rho(\boldsymbol{x}_i))$. We substitute the maximum likelihood estimators $\hat{a}$ and $\hat{b}$ for the a and b in the right-hand side of (5), and we denote the substituted right-hand side by $\hat{\rho}(\boldsymbol{x})$ as an evaluated local defect rate.

From this, we can provide the value of the defect rate by evaluating the multiple integration in (4) numerically, but it is almost impossible if the dimension of $\boldsymbol{x}$ is large. Then, by simulating $\{\tilde{\boldsymbol{x}}_i \mid i = 1, 2, \ldots, \tilde{n}\}$ from the distribution $f(\boldsymbol{x})$ at random, we consider to use Monte Carlo integration, that is, to provide $\sum_{i=1}^{\tilde{n}} \hat{\rho}(\tilde{\boldsymbol{x}}_i)/\tilde{n}$. However, a problem remains. The defect rate is tiny in general, i.e., $\hat{\rho}(\boldsymbol{x}) \approx 0$ for almost all $\boldsymbol{x}$, and so we cannot provide an accurate evaluation of the defect rate even if we simulate huge size of $\{\tilde{\boldsymbol{x}}_i \mid i = 1, 2, \ldots, \tilde{n}\}$.

To overcome this difficulty, we try to simulate $\{\tilde{\boldsymbol{x}}_i \mid i = 1, 2, \ldots, \tilde{n}\}$ from the region where $\hat{\rho}(\boldsymbol{x})$ is large, and then we evaluate the defect rate efficiently by an importance sampling. Concretely speaking, letting $\hat{\boldsymbol{\mu}}$ and $\hat{\Sigma}$ be respectively the sample mean vector and sample variance-covariance matrix for a set of the inputs with a defect $\{\boldsymbol{x}_i \mid y_i = -1\}$, we simulate $\{\tilde{\boldsymbol{x}}_i \mid i = 1, 2, \ldots, \tilde{n}\}$ from the Gaussian distribution $\mathrm{N}(\hat{\boldsymbol{\mu}}, \hat{\Sigma})$ at random. Then we evaluate the defect rate by

$$\frac{1}{\tilde{n}} \sum_{i=1}^{\tilde{n}} \frac{f(\tilde{\boldsymbol{x}}_i)}{g(\tilde{\boldsymbol{x}}_i)} \hat{\rho}(\tilde{\boldsymbol{x}}_i), \tag{6}$$

where $g(\boldsymbol{x})$ is the probability density function of $\mathrm{N}(\hat{\boldsymbol{\mu}}, \hat{\Sigma})$. From the law of large numbers, this converges to our desired expectation in (4).

For the data treated in Section 6, we conducted this defect rate evaluation after 100 iterations of samplings. The estimates of a and b were -3.88 and 0.58, respectively. In Fig. 3, we can check the transition of the evaluations in (6) caused by increasing $\tilde{n}$. The evaluations become stable when $\tilde{n}$ is close to 10^6, and as a result we found that the defect rate is about 1.2×10^{-6}.

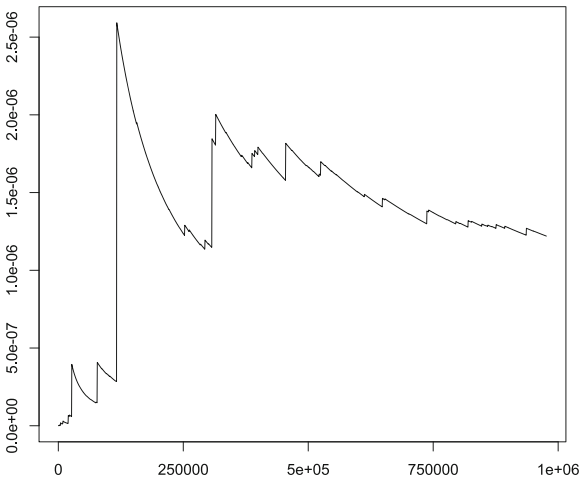

Fig. 3 Transition of estimates for the defect rate. The horizontal axis indicates $\tilde{n}$ and the vertical axis indicates the evaluation in (6)

8 Concluding remarks

In this paper, under the situation where various variables may cause a defect, we have treated a problem to actively estimate the discriminant function which determines the probability of causing the defect. And then, we have discovered that even the ASVM, the latest active learning method in the nonlinear discriminant analysis, does not work well for the case where the nonlinearity of the discriminant function is strong and the region producing the defect more likely than not is separated. To overcome this difficulty, we have proposed the LSVM which uses a linear discriminant method by consciously forgetting the nonlinearity of the discriminant function at the sampling stage in active learning. In numerical studies, we have simulated the cases where the region is actually separated, and then it has been checked that the LSVM is superior to the ASVM for such cases. Also it has been checked through real data analysis that the error rate for the LSVM is smaller than that for the ASVM and becomes stable quickly. Moreover, we have proposed a method to efficiently estimate the defect rate by use of the importance sampling after obtaining the estimated discriminant function by the LSVM. We have used a single Gaussian distribution for the importance sampling, but we may be able to evaluate it faster by using a multi-modal distribution such as a Gaussian mixture.

The above-mentioned case is natural for the defect rate evaluation problem, and so we can say that our simple active learning method is useful in product manufacturing, that is, valuable from engineering viewpoint. On the other hand, brushing up the method is our important future theme in order to cope with the case of existing more variables. One idea is to make a hybrid-type active learning method by combining the LSVM and the ASVM so that the weak and strong points of the ASVM are respectively overcome and kept.

Acknowledgements

The authors would like to thank the reviewer for his/her valuable comments and advice to improve the paper. This research was partially supported by a Grant-in-Aid for Scientific Research (23500353) from the Ministry of Education, Culture, Sports, Science and Technology of Japan.

Author details
[1]Graduate School of Mathematics, Kyushu University, 744, Motooka, Nishi-ku, Fukuoka 819-0395, Japan. [2]Fujitsu Laboratories Ltd, 4-1-1 Kamikodanaka, Nakahara-ku, Kawasaki 211-8588, Japan. [3]Institute of Mathematics for Industry, 744, Motooka, Nishi-ku, Fukuoka 819-0395, Japan.

Received: 28 July 2015 Revised: 21 September 2015
Accepted: 10 October 2015
Published online: 20 October 2015

References

1. Boyd, S, Vandenberghe, L: Convex optimization. Cambridge university press, New York (2009)
2. Cristianini, N, Shawe-Taylor, J: An introduction to support vector machines and other kernel-based learning methods. Cambridge university press, New York (2000)
3. Katayama, K, Hagiwara, S, Tsutsui, H, Ochi, H, Sato, T: Sequential importance sampling for low-probability and high-dimensional SRAM yield analysis. In: Proceedings of the International Conference on Computer-Aided Design, pp. 703–708. IEEE Press, San Jose, California, (2010)
4. Kiefer, J: Optimum experimental designs. J. R. Stat. Soc. Ser. B. **21**, 272–319 (1959)
5. Kiefer, J: Optimum designs in regression problems, II. Ann. Math. Stat. **32**, 298–325 (1961)
6. Kiefer, J, Wolfowitz, J: Optimum designs in regression problems. Ann. Math. Stat. **30**, 271–294 (1959)
7. Platt, J: Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Adv. Large Margin Classifiers. **10**, 61–74 (1999)
8. Rasmussen, CE, Williams, CKI: Gaussian processes for machine learning. MIT Press, Cambridge, MA (2005)
9. Scholköpf, B, Smola, AJ: Learning with kernels: support vector machines, regularization, optimization, and beyond. MIT press, Cambridge, MA (2001)
10. Shawe-Taylor, J, Cristianini, N: Kernel methods for pattern analysis. Cambridge university press, New York (2004)
11. Sun, S, Li, X: Fast statistical analysis of rare circuit failure events via subset simulation in high-dimensional variation space. In: Proceedings of the International Conference on Computer-Aided Design, pp. 324–331. IEEE Press, San Jose, California, (2014)
12. Tong, S, Koller, D: Support vector machine active learning with applications to text classification. J. Mach. Learn. Res. **2**, 45–66 (2002)
13. Umezu, Y, Ninomiya, Y: Optimal experimental design based on Gaussian process classification (in Japanese). In: Proceedings of the Japanese Joint Statistical Meeting, pp. 8–11. University of Osaka, Japan, (2013)