Information Sciences 249 (2013) 85–95

Contents lists available at SciVerse ScienceDirect

Information Sciences

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

The strong convergence of visual classification method and its applications

Deyu Meng [a,*], Yee Leung [b], Zongben Xu [a]

[a] Institute for Information and System Sciences and Ministry of Education Key Lab for Intelligent Networks and Network Security, Xi'an Jiaotong University, Xi'an 710049, PR China
[b] Department of Geography & Resource Management, The Chinese University of Hong Kong,Hong Kong, PR China

ARTICLE INFO

Article history:
Received 7 November 2010
Received in revised form 13 March 2012
Accepted 10 June 2013
Available online 22 June 2013

Keywords:
Disease diagnosis
Pattern classification
Remote sensing
Visual classification method

ABSTRACT

Visual classification method has been proposed as a learning strategy for pattern classification problem. In this paper, we show the strong convergence property of this method. In particular, the method is shown to converge to the Bayesian estimator, i.e., the learning error of the method is convergent to the posterior expected minimal value. The performance of the method has also been theoretically evaluated to comply with the human visual sensation and perception principle. The method is successfully used to some practical remote sensing and disease diagnosis applications. The experimental results all verify the validity and effectiveness of the theoretical conclusions.

1. Introduction

Pattern classification is one of the fundamental problems in pattern recognition. It aims at finding a discriminant rule from a set of experiential data with multiple classes generated from an unknown but fixed distribution, and employing it to categorize any new-coming input datum. Pattern classification has attracted much attention in recent decades due to its wide-spread applications in human, engineering and medical sciences [1,10,24].

Visual classification method (VCM) is one of the latest methods for pattern classification [12,26]. The method is constructed by simulating the human sensation and perception principle. It is capable of implementing effective heuristic pattern categorization similar to the mechanism of human eyes to a certain extent. The main aim of this research is to further propose the theoretical convergence property of the VCM, and make applications in remote sensing and disease diagnosis. In particular, it is proved that the classification discriminant achieved by the VCM is convergent to the Bayesian estimator. That is, the learning error of the VCM tends to attain the Bayesian error. This strong convergence property of the VCM is superior to other pattern classification methods, such as the well-known support vector classification (SVC [18]), which only ensures the convergence of the learning error of the obtained result to the minimal value of a pre-specified learning machine (i.e., a function set). The performance of the VCM in remote sensing and disease diagnosis problems confirm the theoretical conclusions.

In what follows, the general mathematical formulation of the classification problem and a short review of the VCM are first made in Section 2. The theoretical results on the convergence property of the VCM are then proposed in Section 3. The simulation results and the applications in remote sensing and disease diagnosis are discussed in Section 4. We finish with the conclusion in Sections 5.

[*] Corresponding author.
E-mail addresses: dymeng@mail.xjtu.edu.cn (D. Meng), yeeleung@cuhk.edu.hk (Y. Leung), zbxu@mail.xjtu.edu.cn (Z. Xu).

http://dx.doi.org/10.1016/j.ins.2013.06.028

2. Visual classification method

We first give the general mathematical formulation of the classification problem. Since a multi-classification problem can be transformed into a series of two-classification problems, it is generally sufficient to discuss the two-classification problem. Let $D_l = \{x_i, y_i\}_{i=1}^{l}$ be the given two-label training data set, i.i.d. generated from an unknown but fixed distribution $F(x,y) = F(y|x)F(x)$ defined on Z, where $Z = X \times Y, X \subseteq R^n$ is the input (attribute) space, and $Y = \{0,1\}$ is the output (label) space. Given a family of preset indicator functions (i.e., the learning machine) $\mathcal{F} = \{f_\sigma(x), \sigma \in \Lambda\}$, the learning problem aims to select an appropriate discriminant function f_{σ^*} from $\mathcal{F}$ based on the training set D_l so that f_{σ^*} so selected can well implement the underlying classification task. The classification capability of f_{σ^*} can be quantitatively evaluated in a mathematical way as follows.

The *loss function* $L(y_1,y_2)$ $(y_1,y_2 \subset Y)$ is defined as:

$$L(y_1, y_2) = \begin{cases} 0, & \text{if } y_1 = y_2, \\ 1, & \text{if } y_1 \neq y_2. \end{cases} \tag{1}$$

The *risk functional (or risk) of* $f_\sigma \in \mathcal{F}$ is defined as

$$R(f_\sigma) = \int_Z L(y, f_\sigma(x)) dF(x,y) = \int_Z |y - f_\sigma(x)| dF(x,y), \tag{2}$$

which is the expectation of $L(y,f_\sigma(x))$ over Z. A discriminant function f_{σ^*} in $\mathcal{F}$ is of the optimal classification capability if $R(f_\sigma)$ attains its minimum at $\sigma = \sigma^*$ over the entire learning machine $\mathcal{F}$. In these terms, the learning problem can be precisely defined as: finding the optimal discriminant function f_{σ^*} in $\mathcal{F}$ such that

$$R(f_{\sigma^*}) = \min\{R(f_\sigma) : f_\sigma \in \mathcal{F}\} := OPT_F(\mathcal{F}). \tag{3}$$

The quantity $OPT_F(\mathcal{F})$ is called the *minimal risk of the learning machine* $\mathcal{F}$ (with respect to F). Any implementation scheme aiming to find (or approximate) the optimal discriminant function of $\mathcal{F}$ is called *a learning strategy*. Since a learning strategy L is designed on the basis of the given training data D_l, it can thus be viewed as a mapping from the sample set $\mathcal{D}_l$ into the learning machine $\mathcal{F}$. A learning strategy L is *a learning algorithm* if for any $\varepsilon \in (0,1)$ and $\delta \in (0,1)$, there exists an integer $l_0(\varepsilon,\delta)$ such that for any $l > l_0(\varepsilon,\delta)$, it holds that

$$P\{R(L(\mathcal{D}_l)) < OPT_F(\mathcal{F}) + \varepsilon\} \geqslant 1 - \delta, \tag{4}$$

where $L(\mathcal{D}_l)$ is the discriminant function generated from the learning strategy L. In this case, we also say that the learning strategy is convergent. For instance, the SVC is one of the typical convergent learning strategies [18].

It should be noted that $OPT_F(\mathcal{F})$, as defined in (3), is not the essential minimal risk of all nontrivial discriminant functions. The real one is the Bayesian risk, i.e.,

$$OPT_F = \min\{R(f) : f \in \Sigma\},$$

where Σ denotes the collection of all Lebesgue measurable indicator functions defined on X. The Bayesian risk is an intrinsic quantity underlying the learning problem, irrespective of the given learning machine, and no larger than $OPT_F(\mathcal{F})$ for any nontrivial learning machine $\mathcal{F}$. Correspondingly, a learning strategy L is *strongly convergent* if the estimation in (4) holds for the Bayesian risk OPT_F instead of the minimal risk $OPT_F(\mathcal{F})$. Evidently, such a strategy is of better convergence than the aforementioned convergent learning strategy, and hence is always the expected one in real applications.

Accordingly, the learning machine and the learning strategy play a significant role in the final success of pattern classification. Intrinsically speaking, the learning machine utilized in the VCM can be expressed as [26]:

$$\mathcal{F}_{VCM} = \left\{ f_{\sigma,\mathcal{D}_l}(x) = sgn\left(\frac{1}{l} \sum_{i=1}^{l} y_i g(x - x_i, \sigma) \right) : \sigma \geqslant 0 \right\}. \tag{5}$$

Actually, this learning machine can be formulated under the framework of the scale space theory and understood by the visual sensation and perception principle: given a primary image $f(x)$ at the distance σ, the observed blurry image $f(x,\sigma)$ can be mathematically expressed by the following partial differential equation [3,23]:

$$\begin{cases} \frac{\partial f(x,\sigma)}{\partial \sigma} = \frac{\partial^2 f(x,\sigma)}{\partial x^2} \\ f(x,0) = f(x) \end{cases}. \tag{6}$$

The solution of the above equation can be explicitly expressed as

$$f(x,\sigma) = f(x) * g(x,\sigma) = \int g(x-y) f(y) dy,$$

where '$*$' denotes the convolution operation and $g(x,\sigma)$ the Gaussian function

$$g(x,\sigma) = \frac{1}{\left(\sqrt{2\pi}\sigma\right)^n} e^{-\|x\|^2/2\sigma^2}. \tag{7}$$

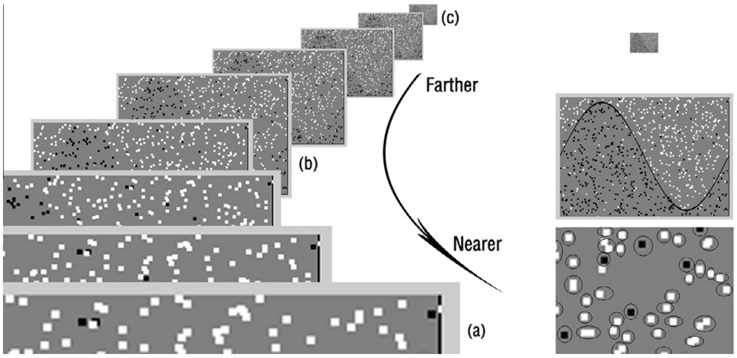

Fig. 1. (a) Observing the data very closely, a discriminant function which exactly separates each point, including involved noises and outliers, can be perceived; (b) Observing the data from a proper distance, the intrinsic discriminant function can be appropriately perceived; and (c) Observing the data from far away, all data are mixed up and no discriminant function can be perceived.

If the training samples are treated as an imaginary image with expression:

$$f(x,\mathcal{D}_l)=\frac{1}{l}\sum_{i=1}^{l}y_i\delta(x-x_i), \tag{8}$$

then the corresponding blurred image at scale σ can be specified as

$$f(x,\sigma,\mathcal{D}_l)=f(x,\mathcal{D}_l)*g(x,\sigma)=\frac{1}{l}\sum_{i=1}^{l}y_ig(x-x_i,\sigma). \tag{9}$$

The learning machine $\mathcal{F}_{VCM}$ as (5) can then be obtained. In fact, the performance of the discriminant function $f_{\sigma,\mathcal{D}_l}(x)$ under different scales σ highly accords with the visual phenomenon of observing the classification image by varying its observing distances, as illustrated in Fig. 1.

Then the learning strategy of the VCM can be easily constructed on the basis of the learning machine $\mathcal{F}_{VCM}$. In fact, any cross validation method, e.g., the k-fold cross-validation method, can be employed [2,17,25] for this task. In this method, the given data set is first partitioned into k subsets. Among these k sets, a single subset is taken as the testing data for measuring the learning error of the related discriminant function, and the remaining $k-1$ sets are utilized as the training data. The cross-validation process is then repeated k times, by taking each of the k subsets as the testing data, respectively. The k results are then averaged to get a single learning error. The appropriate candidates for the scale parameter can then be properly specified through minimizing the cross-validation learning error.

The above strategy defines a mapping $\mathcal{L}_{VCM}$ from the input data $\mathcal{D}_l$ to the learning machine $\mathcal{F}_{VCM}$. We will show in the next section that the mapping $\mathcal{L}_{VCM}$ so defined achieves a strongly convergent learning algorithm by properly specifying the scale parameter.

Remark 1. Actually, by applying the Parzen window method [7] to estimate the densities underlying the positive data and the negative data, respectively, and by comparing the estimated densities at each input datum, a classification discriminant can then be obtained. When the method adopts the Gaussian window, the obtained discriminant is very similar to the discriminant function generated from the VCM. We introduce the VCM in the visual perspective since this will make the following theoretical conclusions more natural and understandable.

3. Theoretical conclusions on the VCM

In this section, we will show that the learning strategy $\mathcal{L}_{VCM}$ converges to the Bayesian estimator. Before presenting the main theorem, we first distinguish two notations of probability.

Denote $\Omega = (X\times Y)$ as the input data space of the classification problem determined by the unknown but fixed distribution $F(x,y)=F(x)F(y|x)$, $\mathbb{F}=M(X)\times S(Y)$ as the σ-algebra of Ω defined by its power set, and P as the probability defined by

$$P(A\times B)=\int_{A\times B}dF(x,y),\quad A\times B\in\mathbb{F}.$$

Then, $\mathcal{P}_1=(\Omega,\mathbb{F},P)$ constitutes a probability space [6]. Likewise, let $\Omega_X = X$ be the attribute space, $\mathbb{F}_X=M(X)$ be the σ-algebra of Ω_X conducted by all the measurable subset X, and the probability P_X be defined by

$$P_X(A)=\int_A dF(x),\quad A\in M(X),$$

and $\mathcal{P}_2 = (\Omega_X, \mathbb{F}_X, P_X)$ also composes of a probability space.

For any function $f_{\sigma,\mathcal{D}_l}(x)$ in $\mathcal{F}_{VCM}$, the upper bound for the deviation of its learning risk from the Bayesian risk can be theoretically estimated, as stated in the following theorem.

Theorem 1. *Let $\mathcal{P}_1 = (\Omega, \mathbb{F}, P)$ and $\mathcal{P}_2 = (\Omega_X, \mathbb{F}_X, P_X)$ be the probability spaces, D_l be the training sample set generated from P, $E_y(x)$ be the average of y at x, and p(x) be the density function of x. Assume that X is open and bounded in R^n and $E_y(x)P(x)$ is continuous on $\overline{X}$ (the closure of X). Then for any fixed $\sigma > 0$, $\delta \in (0,1)$, and $\varepsilon > 0$, there exist positive constants c_1, c_2, c_3, c_4, independent of l and σ, such that*

$$P\{|R(f_{\sigma,\mathcal{D}_l}) - OPT_F| < \varepsilon + P_X\{0 < |E_y(x)p(x)| < Bound(\varepsilon, \delta, l, \sigma)\}\} > 1 - \delta$$

where $Bound(\varepsilon, \delta, l, \sigma)$ is of the form:

$$Bound(\varepsilon, \delta, l, \sigma) = 2\varepsilon + \frac{c_1}{l^{\frac{1}{2}}} + c_2\sigma^{n+2} + c_3\sigma + \frac{c_4}{l^{\frac{1}{2}}(\sigma)^n}. \tag{10}$$

Proof. We only list the main steps of the proof due to the page limitation. The entire proof can refer to the Supplementary material of this paper.

Step 1: If $D_l = \{x_i, y_i\}_{i=1}^l$ is generated from P, $g(x,\sigma)$ is the Gaussian function defined in (7), then for any $z \in R^n$ and $\delta \in (0,1)$, it holds that:

$$P\left(\left|\left(\frac{1}{l}\sum_{i=1}^{l} y_i g(z - x_i, \sigma) - \int yg(z - x, \sigma)dF(x,y)\right)\right| \geqslant \sqrt{\frac{2\ln\frac{2}{\delta}}{l\left(\sqrt{2\pi}\sigma\right)^{2n}}}\right) < \delta.$$

Step 2: For any $\varepsilon > 0$, there exits a constant $\sigma_\varepsilon > 0$ such that for any $\sigma \in (0, \sigma_\varepsilon]$ and $z \in X$, it holds that

$$\left|\int yg(z - x, \sigma)dF(x,y) - E_y(z)P(z)\int_X g(z - x, \sigma)dx\right| < \varepsilon.$$

Step 3: For any σ_1, $\sigma_2 > 0$, it holds that

$$\left|\int yg(z-x,\sigma_1)dF(x,y) - \int yg(z-x,\sigma_2)dF(x,y)\right| < \frac{1}{(2\pi)^{\frac{n}{2}}\min\{\sigma_1,\sigma_2\}^{(n+1)}}\left(n + \frac{2B^2}{\min\{\sigma_1,\sigma_2\}^2}\right)\max\{\sigma_1,\sigma_2\}, \quad \forall z \in X.$$

Step 4: For any $\delta \in (0,1)$ and $\varepsilon > 0$, it holds that

$$P\left\{\left|\left(\frac{1}{l}\sum_{i=1}^{l} y_i g(z - x_i, \sigma) - E_y(z)P(z)A_\sigma(z)\right)\right| \geqslant \frac{a(\delta)}{l^{\frac{1}{2}}(\sigma)^n} + b(\varepsilon)\sigma + \varepsilon = \Psi(\varepsilon, \delta, \sigma, l)\right\} < \delta, \quad \forall \sigma > 0, \quad \forall z \in X,$$

where

$$A_\sigma(z) = \int_X g(z - x, \sigma)dx, \quad a(\delta) = \sqrt{\frac{2\ln\frac{2}{\delta}}{\left(\sqrt{2\pi}\right)^{2n}}}, \quad b(\varepsilon) = \frac{1}{(2\pi)^{\frac{n}{2}}\sigma_\varepsilon^{(n+1)}}\left(n + \frac{2B^2}{\sigma_\varepsilon^2}\right).$$

Step 5: Let

$$M(x) = \begin{cases} E_y(x)P(x)A_\sigma(x), & x \in X \quad \text{and} \quad E_y(x)p(x) \neq 0 \\ f(x, \sigma, \mathcal{D}_l), & x \in X \quad \text{and} \quad E_y(x)p(x) = 0 \end{cases},$$

where $f(x, \sigma, \mathcal{D}_l) = \frac{1}{l}\sum_{i=1}^{l} y_i g(x - x_i, \sigma)$. Then $R(sgn(M(x))) = OPT_F$.

Step 6: Let $f_{\sigma,\mathcal{D}_l}(x) = sgn(f(x, \sigma, D_l))$. Then for any $\sigma > 0, a > 0$ and $z \in X$, we have

$$|R(f_{\sigma,\mathcal{D}_l}) - OPT_F| \leqslant P_X\{|f(x, \sigma, \mathcal{D}_l) - E_y(x)P(x)A_\sigma(x)| \geqslant a)\} + P_X\{0 < |E_y(x)P(x)A_\sigma(x)| < a\}.$$

Step 7: It holds that

$$P\left\{P_X\left\{|f(x, \sigma, \mathcal{D}_l) - E_y(x)P(x)A_\sigma(x)| \geqslant \Psi\left(\frac{\varepsilon}{2}, \frac{\varepsilon\delta}{2}, \sigma, l\right)\right\} \geqslant \frac{\varepsilon}{2}\right\} < \delta.$$

Step 8: It holds that

$$P_X\left\{0 < |E_y(x)P(x)A_\sigma(z)| < \Psi\left(\varepsilon, \frac{\varepsilon\delta}{2}, \sigma, l\right)\right\} < \frac{\varepsilon}{2} + P_X\{0 < |E_y(x)P(x)| < Bound(\varepsilon, \delta, \sigma, l)\},$$

where

$$Bound(\varepsilon,\delta,l,\sigma) = 2\varepsilon + \frac{c_1}{l^{\frac{1}{2}}} + c_2\sigma^{n+2} + c_3\sigma + \frac{c_4}{l^{\frac{1}{2}}(\sigma)^n},$$

where $c_1 = \frac{(2\pi)^{\frac{n}{2}}a\left(\frac{\varepsilon\delta}{2}\right)}{e^{-\frac{2B^2}{\sigma'_\varepsilon}}m(X)}$, $c_2 = \frac{(2\pi)^{\frac{n}{2}}b(\varepsilon)/\sigma'_\varepsilon+(2\pi)^{\frac{n}{2}}\varepsilon/\sigma'^2_\varepsilon}{e^{-\frac{2B^2}{\sigma'_\varepsilon}}m(X)}$, $c_3 = 2b(\varepsilon), c_4 = 2a\left(\frac{\varepsilon\delta}{2}\right)$. Here $\sigma'_\varepsilon > 0$ is a constant independent of l and σ, and $m(X)$ is the m-measure of X.

Step 9: It holds that

$$P\{|R(f_{\sigma,\mathcal{D}_l}) - OPT_F| < \varepsilon + P_X\{0 < |E_y(x)P(x)A_\sigma(x)| < Bound(\varepsilon,\delta,l,\sigma)\}\} > 1-\delta. \qquad \square$$

The significance of Theorem 1 lies on that it provides the following upper bound estimation of the deviation of the learning function risk from the Bayesian risk in probability:

$$R(f_{\sigma,\mathcal{D}_l}) - OPT_F \leqslant \varepsilon + P_X\{0 < |E_y(x)P(x)| < Bound(\varepsilon,\delta,l,\sigma)\}, \tag{11}$$

which also provides a measurement on the generalization capability of any discriminant function $f_{\sigma,\mathcal{D}_l}$ in $\mathcal{F}_{VCM}$.

Formula (11) implies that the task of finding the optimal $f_{\sigma,\mathcal{D}_l}$ from $\mathcal{F}_{VCM}$, where the minimal learning risk is attained, can be easily realized by calculating the minimum of the upper bound $Bound(\varepsilon,\delta,l,\sigma)$. This naturally leads to the following strategy for proper scale parameter selection on $\mathcal{F}_{VCM}$.

Since ε and δ in (10) are arbitrarily valued, $Bound(\varepsilon,\delta,l,\sigma)$ can be taken as a function with respect to l and σ. Some useful observations can then be made:

- $Bound(\varepsilon,\delta,l,\sigma) \to \infty$ when l is fixed and $\sigma \to 0$, since $\frac{c_4}{l^{\frac{1}{2}}(\sigma)^n} \to \infty$ in (10);
- $Bound(\varepsilon,\delta,l,\sigma) \to \infty$ when l is fixed and $\sigma \to \infty$, since $c_2\sigma^{n+2} + c_3\sigma \to \infty$ in (10).

These observations show that whenever σ is too large or too small, $f_{\sigma,\mathcal{D}_l}$ cannot attain good generalization performance. The good performance of the discriminant from the VCM can only occur when the scale parameter σ is set as a moderate value. This fully complies with human visual sensation and perception principle [4,8], as depicted in Fig. 1. This phenomenon also tallies with the conclusion of the traditional statistical learning theory: For too small σ, the classification discriminant tends to be overfitted to the training samples while cannot predict well on new samples; contrarily, for too large σ, the discriminant tends to be configured very smoothly while cannot well fit input samples. Only at a moderate value of σ, a good compromise between empirical risk and generalization error of the VCM is expected to be achieved [19,20].

There remains another problem: Is there an optimal scale σ^* and where is it? The following theorem provides an answer to this problem:

Theorem 2. *For any fixed l, the function Bound($\varepsilon,\delta,l,\sigma$) is of the unique minimum, attained at:*

$$\sigma^* = Cl^{-\frac{1}{2n+2}}, \tag{12}$$

where n is the dimension of the attribute space and

$$C = \sqrt[n+1]{\frac{2nc_4}{c_3 + \sqrt{c_3^2 + 4(n+2)c_2\frac{nc_4}{l^{\frac{1}{2}}}}}}, \tag{13}$$

where c_1, c_2, c_3, c_4 are the constants as defined in Theorem 1.

Proof. From (10), it follows that

$$f(\sigma) = \frac{\partial Bound(\varepsilon,\delta,l,\sigma)}{\partial\sigma} = (n+2)c_2\sigma^{n+1} + c_3 - \frac{nc_4}{l^{\frac{1}{2}}(\sigma)^{n+1}} = \frac{1}{(\sigma)^{n+1}}\left((n+2)c_2(\sigma^{n+1})^2 + c_3\sigma^{n+1} - \frac{nc_4}{l^{\frac{1}{2}}}\right). \tag{14}$$

It is easy to see that there are two solutions for $f(\sigma) = 0$, and one is positive and the other negative. Denote the positive solution as σ^*. Then $f(\sigma) < 0$ whenever $0 < \sigma < \sigma^*$ and $f(\sigma) > 0$ whenever $\sigma > \sigma^*$ since $f(0)$ is negative and $f(+\infty)$ is positive. Consequently σ^* is the unique minimum of $Bound(\varepsilon,\delta,l,\sigma)$ on $[0,+\infty)$. From (14), we have

$$\sigma^* = \left(\frac{-c_3 + \sqrt{c_3^2 + 4(n+2)c_2\frac{nc_4}{l^{\frac{1}{2}}}}}{2(n+2)c_2}\right)^{\frac{1}{n+1}} = \left(\frac{1}{c_3 + \sqrt{c_3^2 + 4(n+2)c_2\frac{nc_4}{l^{\frac{1}{2}}}}}\frac{4(n+2)c_2\frac{nc_4}{l^{\frac{1}{2}}}}{2(n+2)c_2}\right)^{\frac{1}{n+1}}$$

$$= \left(\frac{2nc_4}{c_3 + \sqrt{c_3^2 + 4(n+2)c_2\frac{nc_4}{l^{\frac{1}{2}}}}}l^{-\frac{1}{2}}\right)^{\frac{1}{n+1}}.$$

The proof of Theorem 2 is completed. □

The above theorem implies that the preferred scale parameter of the VCM should be with the rank $O\left(l^{-\frac{1}{2n+2}}\right)$. That is to say, we can set the selected range of the scale as follows: first, specify the appropriate parameters a and b (say, $a = 10^{-1}$ and $b = 10^2$), and next, let $\varepsilon_l = al^{-\frac{1}{2n+2}}$ and $E_l = bl^{-\frac{1}{2n+2}}$, and then, select the optimal scale in $[\varepsilon_l, E_l]$.

According to the above discussion, the mapping $\mathcal{L}_{VCM} : \bigcup \mathcal{D}_l \to \mathcal{F}_{VCM}$ can be formulated as

$$L_{VCM}(\mathcal{D}_l) = f_{\sigma^*,\mathcal{D}_l}(x), \tag{15}$$

where σ^* is calculated by the following optimization:

$$\sigma^* = \arg\min_{\sigma\in[\varepsilon_l,E_l]} CV(\sigma), \tag{16}$$

where $CV(\sigma)$ is the cross validation error under the scale parameter σ. To minimize the continuous function $CV(\sigma)$ and find the optimal scale parameter σ^* in (16), any one-dimensional global optimization method, such as the grid algorithm [15], the simulated annealing [11] and the evolutionary method [14], can be employed.

The following theorem shows that the above learning strategy $\mathcal{L}_{VCM}$ is a strongly convergent learning algorithm.

Theorem 3. *Assume that the attribute space X is open and bounded in R^n and $E_y(x)P(x)$ is continuous on $\overline{X}$ (the closure of X), then for any $\varepsilon > 0$ and $\delta \in (0,1)$, there exists $l(\varepsilon,\delta)$ such that for any $l > l(\varepsilon,\delta)$, it holds that*

$$P\{R(L_{VCM}(\mathcal{D}_l) - OPT_F) \geqslant \varepsilon\} < \delta. \tag{17}$$

Proof. We prove Theorem 3 by the following four steps. The entire proof can refer to the Supplementary material of this paper.

Step 1: Suppose X is an m-measurable subset in R^n, and $\{f_n\}$ is a sequence of measurable functions satisfying $f_n(x) \to f(x)$, $x \in X$, *as* $n \to \infty$. If there exists an m-measurable function g on X such that $|f_n(x)| \leqslant g(x)$, $n = 1, 2, \ldots, x \in X$, then

$$\lim_{n\to\infty} \int_X f_n(x)dx = \int_X f(x)dx.$$

Step 2: It holds that $\lim_{\varepsilon\to 0} P_X\{0 < |E_y(x)P(x)| < \varepsilon\} = 0$.
Step 3: It holds that $\lim_{l\to\infty} Bound(\varepsilon,\delta,l,\sigma_l^*) = 2\varepsilon$, where σ_l^* is specified by Formula (16).
Step 4: For any $\varepsilon > 0$ and $\delta \in (0,1)$, there exists an integer $l(\varepsilon, \delta)$ such that for any $l > l(\varepsilon, \delta)$, it holds that

$$P\left\{\left(R\left(f_{\sigma_l^*,\mathcal{D}_l}(x)\right) - OPT_F\right) < \varepsilon\right\} > 1 - \delta. \qquad \square$$

In the next section, we further verify these theoretical conclusions by experimental results.

4. Simulations and applications in remote sensing and disease diagnosis of the VCM

In this section we provide a series of synthetic and real experimental results to support the validity of the presented theoretical results. The optimization problem (16) was solved by the grid optimization method.

4.1. Synthetic simulations

The first set of simulations was designed to demonstrate the validity of using $Bound(\varepsilon,\delta,l,\sigma)$ in Theorem 1 as the upper bound for the deviation of the risk of a discriminant function $f_{\sigma,\mathcal{D}_l}(x) \in \mathcal{F}_{VCM}$ from the Bayesian risk, i.e., the feasibility of applying this estimated bound to quantitatively measure the classification capability of $f_{\sigma,\mathcal{D}_l}(x)$. The simulations were implemented by comparing the behavior of $Bound(\varepsilon,\delta,l,\sigma)$ and the performance of the discriminant function $f_{\sigma,\mathcal{D}_l}(x)$. The spiral classification data $D_{100} = \{x_i^+, +1\}_{i=1}^{50} \cup \{x_i^-, -1\}_{i=1}^{50}$ were utilized, where

$$x_i^+ = (exp((-1.5\pi + i\pi/30)) - 0.5) * cos(-1.5\pi + i\pi/10);$$
$$x_i^- = (exp((-1.5\pi + i\pi/30)) - 0.5) * sin(-1.5\pi + i\pi/10).$$

With the scale parameter σ varying from small to large, the performance of the discriminant function $f_{\sigma,\mathcal{D}_l}$ is demonstrated in Fig. 2. It can be observed that $f_{\sigma,\mathcal{D}_l}$ has very poor performance when σ is too large or too small. Correspondingly, it is easy to deduce that the value of the bound function $Bound(\varepsilon,\delta,l,\sigma)$ varies from infinitely large to its minimum, and then to infinitely large again, as σ goes from 0 to infinity. The observed performance of $f_{\sigma,\mathcal{D}_l}(x)$ clearly accords with the behavior of $Bound(\varepsilon,\delta,l,\sigma)$. This verifies the rationality of the proposed upper bound estimation (Theorem 1).

The second set of simulations was designed to verify the reasonability of Theorem 2, that is, to show that there exists a positive constant C such that the optimal scale σ_l^* and the data size l obey the relation

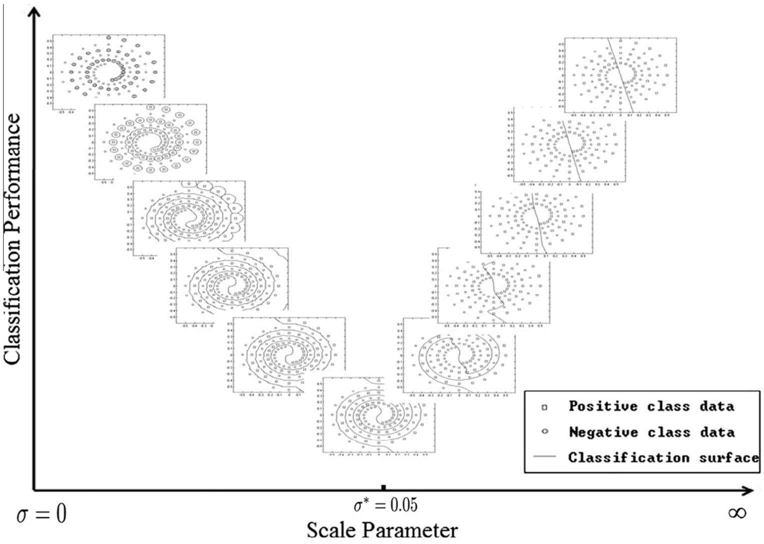

Fig. 2. The classification performance of the discriminant function $f_{\sigma,\mathcal{D}_l}$ with the scale parameter σ varying from very small to very large. σ^* is the optimal scale parameter attained by VCM.

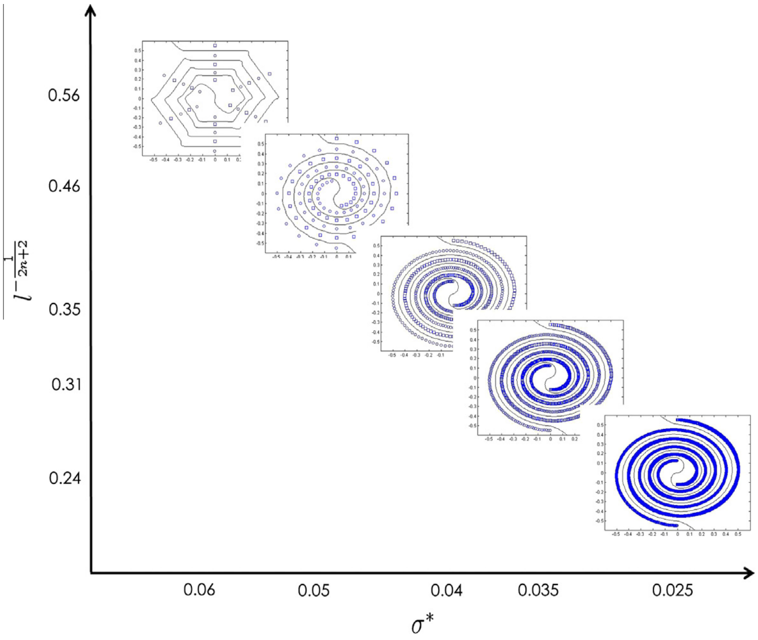

Fig. 3. The classification performance of the optimal discriminant function $f_{\sigma_l^*,\ \mathcal{D}_l}$ obtained by VCM with varying data sizes. It can be observed that all $\left(\sigma_l^*, l^{-\frac{1}{2n+2}}\right)$ are nearly on a line, that is, $l^{-\frac{1}{2n+2}}/\sigma_l^*$ is approximately a constant for any l.

$$C = l^{-\frac{1}{2n+2}}/\sigma_l^*. \tag{18}$$

The spiral data sets $D_l = \{x_i^+, +1\}_{i=1}^{l/2} \cup \{x_i^-, -1\}_{i=1}^{l/2}$ with varying sizes were employed to substantiate (18), where

$$x_i^+ = (exp((-1.5\pi + i\pi/0.3l)) - 0.5) * cos(-1.5\pi + i\pi/0.1l);$$
$$x_i^- = (exp((-1.5\pi + i\pi/0.3l)) - 0.5) * sin(-1.5\pi + i\pi/0.1l).$$

By applying the VCM, the optimal scales σ_l^* are calculated as 0.06, 0.05, 0.04, 0.035, and 0.025 corresponding to l = 30, 100, 500, 1000, and 5000, respectively, as shown in Fig. 3. In these simulations, $C_{30} \approx 0.1057$, $C_{100} \approx 0.1077$, $C_{500} \approx 0.1126$, $C_{1000} \approx 0.1107$ and $C_{5000} \approx 0.1134$. All are approximately equal. This supports the validity of Theorem 2.

Table 1
The misclassification rates (%) of VCM, SVC, LMNN and KNN on the third series of simulation data with respect to varying data sizes. ∗ Means the method is infeasible on the corresponding data.

Method	Training size											
	100	200	400	800	1600	4000	8000	12,000	20,000	30,000	50,000	100,000
VCM	25.46	25.44	24.52	23.58	22.34	21.38	20.60	20.48	20.40	20.38	20.23	20.04
SVC	28.02	25.02	23.88	23.74	22.40	21.78	21.28	21.50	21.24	20.40	20.54	20.42
LMNN	20.96	22.10	21.52	22.12	21.00	21.10	20.94	20.90	∗	∗	∗	∗
KNN	21.00	23.50	21.50	21.12	21.31	21.02	20.92	20.97	20.67	20.80	∗	∗

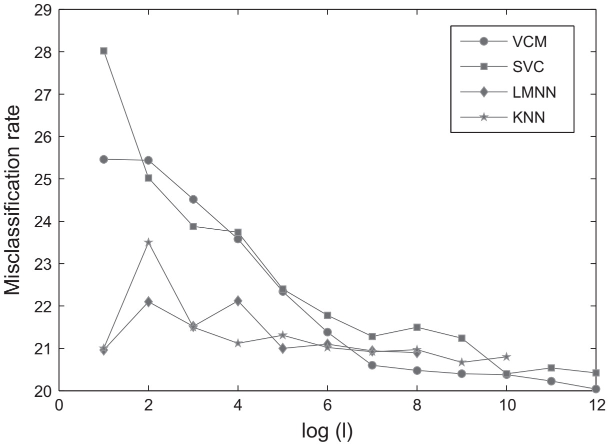

Fig. 4. The tendency curves of the misclassification rates obtained by VCM, SVC, LMNN and KNN on the third series of synthetic simulations.

The third set of simulations was run to demonstrate the strong convergence property of the VCM (namely, Theorem 3). To this aim, we applied the VCM to 12 groups of training data sets $D_l = \left\{\left(x_i^{(1)}, x_i^{(2)}\right), y_i\right\}_{i=1}^{l}$, with varying sizes l = 100, 200, 400, 800, 1600, 4000, 8000, 12,000, 20,000, 30,000, 50,000, 100,000, respectively. All data were generated from the distribution $F((x^{(1)},x^{(2)}),y) = F(y|(x^{(1)},x^{(2)}))F((x^{(1)},x^{(2)}))$ with $F((x^{(1)},x^{(2)}))$ being the uniform distribution on $[0,1] \times [-1,1]$ and

$$F(1|(x^{(1)},x^{(2)})) = \begin{cases} 0.8, & \text{if } x^{(2)} \geqslant 0, \\ 0.2, & \text{if } x^{(2)} < 0; \end{cases} \text{ and } F(-1|(x^{(1)},x^{(2)})) = \begin{cases} 0.2, & \text{if } x^{(2)} \geqslant 0, \\ 0.8, & \text{if } x^{(2)} < 0. \end{cases}$$

The Bayesian risk of this classification problem can be easily calculated as $OPT_F = 0.2$. To evaluate the performance of the VCM, a set of 5000 data, which was i.i.d. generated from the distribution $F((x^{(1)},x^{(2)}),y)$, was employed to compute the misclassification rate of the VCM in each case. For comparison, the support vector classification (SVC [18]), the large margin nearest neighbor method (LMNN [21,22]) and the k-nearest-neighbor classification method (KNN [5]) have also been implemented on all data. The 5-fold cross-validation method [2] was implemented on the corresponding training data for parameter selection in all simulations.

For each of the training sets, the VCM, SVC, LMNN and KNN attained four classifiers, respectively, and the misclassification rates of the classifiers were then evaluated on the testing set. Table 1 lists the misclassification rates of these methods for each training data set, and Fig. 4 shows the tendency curves to depict the convergence of the utilized methods. From Table 1 and Fig. 4, it is easy to see that the VCM monotonically converges to the Bayesian risk 0.2 as the data size l increases, while the SVC, LMNN and KNN do not[1]. This verifies the rationality of Theorem 3.

[1] Two intrinsic properties of the LMNN and KNN can be observed from this simulation. The first is that both methods have a very stable performance with respect to the training sizes. When data size is small, LMNN and KNN evidently outperform other methods, while when the data size becomes larger, the advantage of the VCM tends to gradually arise due to its strong convergence property. The second is that in large data cases, the LMNN and KNN methods tend to be infeasible (when l > 12,000 and l > 30,000, respectively, in this simulation). This is because in LMNN and KNN, a distance matrix between all data pairs needs to be recorded based on the input data. This makes the computational complexity of each method relatively high. As a comparison, the VCM can deal with much larger problems, as discussed in the conclusion of the paper.

Table 2
Information of the mangrove data. NTr2005, NTe2005, NTr2006, NTe2006 denote the numbers of training and testing data sets of the mangrove data collected in 2005 and 2006, respectively.

Class	Data			
	NTr2005	NTe2005	NTr2006	NTe2006
ω_1: Aegiceras corniculatum	215	89	527	222
ω_2: Acanthus ilicifolius	975	324	1370	530
ω_3: Avicennia marina	552	225	607	234
ω_4: Kandelia obovata	1397	522	1090	406
ω_5: Sonneratia caseolaris	918	329	463	198
Total	4057	1489	4057	1590

Table 3
Misclassification rates (%) of VCM, SVC, LMNN and KNN on mangrove data.

Method	Data	
	2005data	2006data
VCM	4.50	7.55
SVC	5.57	8.68
LMNN	4.83	7.50
KNN	5.84	8.92

4.2. Mangrove species identification with high resolution Quickbird images

This empirical study deals with the classification of the mangrove species in remotely sensed images. Two sets of high resolution Quickbird images acquired in 2005 and 2006 were employed to empirically verify the aforementioned convergence property of the VCM. Both sets contain the data of five known mangrove species. The data were collected from the Inner Deep Bay Area of Mai Po (northwestern part of Hong Kong) in 2005 and 2006, respectively. A prime component of the land cover is an extensive mangrove stand with five different mangrove species. Field data were collected and was randomly partitioned for training and testing purposes, respectively. These data were acquired at 4 spectral bands (blue, green, red and near-infrared) at 2.4 m spatial resolution. The basic information is listed in Table 2.

The SVC, LMNN, KNN and VCM were implemented on both data sets, and 5-fold cross-validation was utilized for parameter selection. The experimental results are listed in Table 3. From the table, it is evident that the VCM is of smaller misclassification rates than SVC and KNN on both data sets, and performs similarly good as the LMNN. Thus the theoretical advantage of the VCM can be further substantiated.

4.3. Disease diagnosis

Recently, there arise multiple applications of data mining algorithms in the computer-aided diagnosis [9,13,27]. To further verify the theoretical results on the VCM (especially its strong convergence property), we applied the VCM to some disease diagnosis problems. In particular, four sets of data were utilized, as introduced in the following:

- *Breast cancer data*. This data set was obtained from the University Medical Centre, Institute of Oncology, Ljubljana, Yugoslavia. The instances are described by 9 attributes: age, menopause, tumor-size, inv-nodes, node-caps, deg-malig, breast, breast-quad and irradiat. The output classes are non-recurrence or recurrence of the event.
- *Diabetes disease data*. This data set was obtained from the National Institute of Diabetes and Digestive and Kidney Diseases. There are 8 input attributes: number of times being pregnant, plasma glucose concentration a 2 h in an oral glucose tolerance test, diastolic blood pressure, triceps skin fold thickness, 2-h serum insulin, body mass index, diabetes pedigree function and age. The classes are the non-occurrence and occurrence of the disease.
- *Heart disease data*. This database contains 76 attributes, but all published experiments only need 14 of them. The 13 input attributes used are age, sex, chest pain type, resting blood pressure, serum cholest-oral, fasting blood sugar, resting electrocardiographic results, maximum heart rate achieved, exercise induced angina, oldpeak, the slope of the peak exercise ST segment and number of major vessels. The classes are the absence or presence of heart disease.
- *Thyroid disease data*. This data set was collected by several laboratory tests used to predict whether or not a patient's thyroid belongs to the class euthyroidism, hypothyroidism or hyperthyroidism. The 5 input attributes include: T3-resin uptake test, total Serum thyroxin as measured by the isotopic displacement method, total serum triiodothyronine as measured by radioimmuno assay, basal thyroid-stimulating hormone as measured by radioimmuno assay and maximal absolute difference of TSH value after injection of 200 micro grams of thyrotropin-releasing hormone as compared to the basal value. The 3 output classes represent the euthyroidism, hypothyroidism and hyperthyroidism diagnosis result, respectively.

Table 4
Statistics of the data for the 4 disease diagnosis classification problems used in simulations A1, A2, and A3.

Problems	Dim	Training data sizes (A1,A2,A3)	Testing data sizes (A1,A2,A3)
Breast-cancer	9	200 × 100, 1000 × 20, 2000 × 10	77 × 100, 385 × 20, 770 × 10
Diabetes	8	468 × 100, 2340 × 20, 4680 × 10	300 × 100, 1500 × 20, 3000 × 10
Heart	13	170 × 100, 850 × 20, 1700 × 10	100 × 100, 500 × 20, 1000 × 10
Thyroid	5	140 × 100, 700 × 20, 1400 × 10	75 × 100, 375 × 20, 750 × 10

Table 5
Misclassification rates of SVC, LMNN, KNN and VCM on disease diagnosis data. The best results are highlighted in bold.

Problems	SVC (%)	Misclassification rate (A1)		VCM (%)	Misclassification rate (A2)	
		LMNN (%)	KNN (%)		SVC (%)	LMNN (%)
Breast-cancer	**25.48 ± 4.41**	26.36 ± 0.19	28.27 ± 0.01	25.69 ± 3.38	2.89 ± 0.62	**2.47 ± 0.01**
Diabetes	**23.51 ± 1.48**	25.02 ± 0.02	27.19 ± 0.01	25.84 ± 1.61	0.48 ± 0.54	**0.46 ± 0.02**
Heart	**15.62 ± 3.26**	17.13 ± 0.08	17.88 ± 0.12	17.19 ± 3.00	0.30 ± 0.57	0.32 ± 0.02
Thyroid	5.07 ± 2.33	4.96 ± 0.05	6.40 ± 0.00	**4.28 ± 1.87**	**0.07 ± 0.30**	0.23 ± 0.03
Average	**17.42 ± 2.87**	18.37 ± 0.09	19.94 ± 0.03	18.25 ± 2.47	0.94 ± 0.51	**0.87 ± 0.02**
	KNN (%)	Misclassification rate (A2)		SVC (%)	Misclassification rate (A3)	
		VCM (%)		LMNN (%)	KNN (%)	VCM (%)
Breast-cancer	4.03 ± 0.01	2.84 ± 0.51	2.84 ± 0.34	2.99 ± 0.02	2.99 ± 0.01	**2.84 ± 0.30**
Diabetes	0.48 ± 0.00	0.53 ± 0.56	0.07 ± 0.21	**0.03 ± 0.01**	**0.03 ± 0.01**	0.07 ± 0.12
Heart	**0.28 ± 0.01**	0.45 ± 0.60	**0.0 ± 0.0**	**0.0 ± 0.0**	**0.0 ± 0.0**	**0.0 ± 0.0**
Thyroid	0.64 ± 0.00	0.27 ± 0.82	**0.0 ± 0.0**	**0.0 ± 0.0**	**0.0 ± 0.0**	**0.0 ± 0.0**
Average	1.36 ± 0.01	1.02 ± 0.62	0.73 ± 0.14	0.75 ± 0.01	0.76 ± 0.01	**0.73 ± 0.10**

To reasonably evaluate the classification capability of the utilized learning strategies, the following way is specifically designed: first, multiple partitions of the training and testing sets were randomly generated from the original data sets. Second, on each partition, a classifier was trained on the training set by using the learning strategy and its testing error was obtained on the testing set, correspondingly; the mean and the variance of these testing errors were then taken as the final criterion for evaluating the capability of the strategy. Following this line, we have implemented three series of simulations (denoted as A1, A2 and A3, respectively). In series A1, 100 partitions of the training and testing sets were used, which can be directly downloaded from the website: http://archive.ics.uci.edu/ml/. In series A2 and A3, there are 20 and 10 partitions of the training and testing sets, respectively, formed by combining each 5 and 10 training and testing sets of those used in A1. The statistics of all the utilized data are listed in Table 4. The SVC, LMNN and KNN were applied to these data for comparison. The experimental results are summarized in Table 5. The 5-fold cross-validation method was employed as the parameter selection strategy in all experiments.

From Table 5, it is easy to see that when the training data size increases, the VCM gradually outperforms other utilized methods. Particularly, in A1 simulations, the SVC performs better in 3 data, and the VCM only in 1. In A2 simulations, SVC, LMNN and KNN show advantages in 1, 2, 1 cases, respectively, while the VCM is not the best in all cases. Yet in A3 simulations where the data size is largest, the VCM tends to outperform SVC, LMNN and KNN. In specific, in 3 of 4 applications, the misclassification rates of the VCM is no larger than those of the SVC, LMNN and KNN. Furthermore, the VCM is of the best performance in average. This substantiates the strong convergence property of the VCM in disease diagnosis experiments.

5. Conclusion and discussion

In this paper, we have shown the strong convergence property of the VCM method, i.e., the method converges to the Bayesian solution of the classification problem. The theoretical result has also been substantiated by a series of synthetic simulations and applications in remote sensing and disease diagnosis. Compared with the current classification methods, the VCM has been demonstrated to be effective and efficient.

It should be noted that only simple computations are involved in the implementation of the VCM, and it is easy to conduct that it only needs $O(dl)$ computational cost for training, where d and l denote the dimension and the size of the training data, respectively. That is to say, the computational time of the VCM linearly increases with respect to both the dimension and the size of the input data. Comparatively, the SVC and the LMNN need around $O(dl^{2.2})$ [16] and $O(dl^2)$ [22] costs for training, respectively. Especially, the LMNN needs to learn a Mahanalobis metric by the semidefinite programming technique in the training process. This makes the method always unavailable to large data, as illustrated in the simulations of Section 4.1.

It should also be emphasized that for data sets with small or middle sized training samples, the VCM still cannot always guarantee a better performance than other classification techniques. This phenomenon can be easily observed in cases of the disease diagnosis experiments. How to further improve the capability of the VCM in small-sample cases thus needs to be more investigated in our future research. Besides, further effort needs to be made to investigate the influence of data quality, such as data mixed with noises, outliers and artifacts, to the performance of the VCM theoretically and empirically.

Acknowledgements

This research was supported by the Geographical Modeling and Geocomputation Program under the Focused Investment Scheme at The Chinese University of Hong Kong, the National Grand Fundamental Research 973 Program of China under Grant No. 2013CB329404, the China NSFC Project under Contract 11131006 and Ph.D. Programs Foundation of Ministry of Education of China 20090201120056.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.ins.2013.06.028.

References

[1] U. Alper, M. Alper, C.R. Babu, Mr(2)PSO: a maximum relevance minimum redundancy feature selection method based on swarm intelligence for support vector machine classification, Information Sciences 181 (20) (2011) 4625–4641.
[2] Y. Bengio, Y. Grandvalet, No unbiased estimator of the variance of K-fold cross-validation, Journal of Machine Learning Research 5 (2004) 1089–1105.
[3] J.R. Cannon, The one-dimensional heat equation, Encyclopedia of Mathematics and Its Applications 15 (1984).
[4] S. Coren, L. Ward, J. Enns, Sensation and Perception, Harcourt Brace College Publishers, 1994.
[5] T.M. Cover, P.E. Hart, Nearest neighbor pattern classification, IEEE Transactions on Information Theory 13 (1967) 21–27.
[6] L. Devroye, L. Gyorfi, G. Lugosi, Probabilistic Theory of Pattern Recognition, Springer, 1996.
[7] R.O. Duda, P.E. Hart, D.G. Stork, Pattern Classification, Wiley, New York, 2004.
[8] E.B. Goldstein, Sensation and Perception, Wadsworth Thomson Learning, 2002.
[9] Q.H. Hu, W. Pan, S. An, P.J. Ma, J.M. Wei, An efficient gene selection technique for cancer recognition based on neighborhood mutual information, International Journal of Machine Learning and Cybernetics 1 (2010) 63–74.
[10] Y.C. Hu, C.J. Chen, A PROMETHEE-based classification method using concordance and discordance relations and its application to bankruptcy prediction, Information Sciences 181 (22) (2011) 4959–4968.
[11] T.J. Jose, R.T. Eduardo, New bounds for binary covering arrays using simulated annealing, Information Sciences 185 (1) (2012) 137–152.
[12] Y. Leung, J.S. Zhang, Z.B. Xu, Clustering by scale space filtering, IEEE Transactions on Pattern Analysis and Machine Intelligence 22 (2000) 2–15.
[13] Z. Liu, Q.H. Wu, Y. Zhang, C.L.F. Chen, Adaptive least squares support vector machines filter for hand tremor canceling in microsurgery, International Journal of Machine Learning and Cybernetics 1 (2011) 37–47.
[14] H.N. Luong, H.T. Nguyen, A.C. Wook, Entropy-based efficiency enhancement techniques for evolutionary algorithms, Information Sciences 188 (1) (2012) 100–120.
[15] C. Padgett, K. KreutzDelgado, A grid algorithm for autonomous star identification, IEEE Transaction on Aerospace and Electronic Systems 33 (1) (1997) 202–213.
[16] J. Platt, Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines. Technical Report MSR-TR-98-14, Microsoft Research, 1998.
[17] Y.O. Taner, Model selection in omnivariate decision trees using structural risk minimization, Information Sciences 181 (23) (2011) 5214–5226.
[18] V.N. Vapnik, Statistical Learning Theory, J. Willey, New York, 1998.
[19] X.Z. Wang, C.R. Dong, Improving generalization of fuzzy if–then rules by maximizing fuzzy entropy, IEEE Transactions on Fuzzy Systems 17 (2009) 556–567.
[20] X.Z. Wang, J.H. Zhai, S.X. Lu, Induction of multiple fuzzy decision trees based on rough set technique, Information Sciences 178 (2008) 3188–3202.
[21] K.Q. Weinberger, J. Blizer, L.K. Saul, Distance metric learning for large margin nearest neighbor classification., Advances in Neural Information Processing Systems 18 (2006) 1473–1480.
[22] K.Q. Weinberger, L.K. Saul, Distance metric learning for large margin nearest neighbor classification, Journal of Machine Learning Research 10 (2009) 207–244.
[23] A.P. Witkin, Scale-space filtering, in: Proc. 8th IJCAI, 1983, pp. 1019–1022.
[24] R. Xia, C.Q. Zong, S.S. Li, Ensemble of feature sets and classification algorithms for sentiment classification, Information Sciences 181 (6) (2011) 1138–1152.
[25] Z.B. Xu, M.W. Dai, D.Y. Meng, A fast heuristic strategy for model selection of support vector machines, IEEE Transactions on Systems, Man and Cybernetics, Part B 39 (5) (2009) 1292–1307.
[26] Z.B. Xu, D.Y. Meng, W.F. Jing, A new approach for classification: visual simulation point of view, Lecture Notes in Computer Science 3497 (Part II) (2005) 1–7.
[27] S.M. Zhang, P. McCullagh, C. Nugent, H.R. Zheng, M. Baumgarten, Optimal model selection for posture recognition in home-based healthcare, International Journal of Machine Learning and Cybernetics 1 (2011) 1–14.