The strong convergence of visual classification method and its applications in disease diagnosis

Deyu Meng[1], Zongben Xu[1], Yee Leung[2,*], Tung Fung[2]

[1] Institute for Information and System Science, Xi'an Jiaotong University, Xi'an, 710049,China

[2] Department of Geography & Resource Management, The Chinese University of Hong Kong, Shatin, Hong Kong

yeeleung@cuhk.edu.hk

Abstract. Visual classification method is introduced as a learning strategy for pattern classification problem in bioinformatics. In this paper, we show the strong convergence property of the proposed method. In particular, the method is shown to converge to the Bayes estimator, i.e., the learning error of the method tends to achieve the posterior expected minimal value. The method is successfully applied to some practical disease diagnosis problems. 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 in general and bioinformatics in particular. It aims at finding a discriminant rule from a set of experiential data with multiple labels generated from an unknown but fixed distribution, and then, according to the rule found, categorizing any new input data. Pattern classification has attracted extensive attention in recent decades due to its wide-spread applications in human, engineering and medical sciences. Typical examples are, for instance, handwritten digit recognition[1], face detection[2], bio-sequence analysis[3], structure prediction[4], drug design[5] etc.

Visual classification method (VCM) is one of the latest methods for pattern classification[6, 7]. The method is constructed by mimicking the human sensation and perception principle. It, to a certain extent, can implement effective heuristic categorization of patterns similar to the mechanism of human eyes. The main aim of this research is to further propose the theoretical convergence property of the VCM, and make applications in disease diagnosis. In particular, it is proved that the classification discriminant function obtained by the VCM is convergent to the Bayes estimator. That is, the learning error of the VCM tends to achieve the posterior expected minimal value. This strong convergence property of the

* This work was supported by the earmarked Grant CUHK 4701/06H of the Hong Kong Research Grants Council.

** * is the corresponding author.

M. Chetty, S. Ahmad, A. Ngom, S. W. Teng (Eds.): PRIB2008, Supplementary Conf. Proc., pp.83-94, 2008.

VCM is superior to other pattern classification methods, such as the well-known support vector classification (SVC [8]), which only ensures the convergence of the learning error of the obtained result to the minimal values of a pre-specified learning machine (i.e., a function set). The performances of both methods in disease diagnosis applications all confirm the theoretical conclusions.

In what follows, the general mathematical formulation of the classification problem and a short review of the VCM is first made in Section 2. The theoretical conclusions on the convergence property of the VCM are then proposed in Section 3. Taking the SVC as a basis for comparison, the simulation results and the applications in disease diagnosis are discussed in Section 4. A brief summary of the paper is lastly given in Section 5.

2 Visual Classification Method

To facilitate our discussion, we first give the general mathematical formulation of the classification problem. Since many multiple-label classification problems can be transformed into a series of two-label problems, it is generally sufficient to discuss the two-label classification problem. Let $D_l = \{x_i, y_i\}_{i=1}^{l}$ be a given two-label training data set which is independently 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 $Y = \{-1, +1\}$ is the output (label) space. Given a family of preset indicator functions $\mathcal{F} = \{f_\sigma(x), \sigma \in \Lambda\}$, or equivalently, a learning machine, the learning problem (precisely, the classification learning problem) aims at determining an appropriate function f_{σ^*} from $\mathcal{F}$ by virtue of the training set D_l such that f_{σ^*} has optimal classification capability for the classification problem in a certain sense. The optimal classification capability for the function f_{σ^*} in $\mathcal{F}$ can be mathematically expressed using the following definitions:

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_σ *on* Z is defined as

$$R(f_\sigma) = \int_Z L(y, f_\sigma(x)) dF(x, y) = \frac{1}{2} \int_Z |y - f_\sigma(x)| dF(x, y) \tag{2}$$

which is the expectation value of $L(y, f_\sigma(x))$ over Z. A discriminant function f_{σ^*} in $\mathcal{F}$ has optimal classification capability for the classification problem (or equivalently, f_{σ^*} is an optimal discriminant function in $\mathcal{F}$) if its risk is the minimization of the risks over the learning machine $\mathcal{F}$. In these terms, the learning problem can be more 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 in (3) is often called the *minimal risk of the machine* $\mathcal{F}$ (with respect to F). Any implementation scheme to find (or approximate) the optimal discriminant function of $\mathcal{F}$ is called *a learning strategy*. A learning strategy L is normally designed on the basis of the given training sample set 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 said to be a *learning algorithm* if for any $\varepsilon \in (0,1)$ and any $\delta \in (0,1)$, there is an integer $l_0(\varepsilon,\delta)$ such that whenever $l > l_0(\varepsilon,\delta)$,

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

where $L(\mathcal{D}_l)$ is the discriminant function generated from the learning strategy. In this case, we also say that the learning strategy is convergent. For instance, the SVC is one of the typical convergent learning strategies[8].

It shall 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 collections of all Lebesgue measurable indicator function defined on X. It is easy to know that the Bayesian risk is an intrinsic quantity of a learning problem, irrespective of the given learning machine, and is no larger than $OPT_F(\mathcal{F})$ of any nontrivial learning machine $\mathcal{F}$. Correspondingly, a learning strategy L is said to be *strongly convergent* when the estimation in (4) holds for the Bayesian risk OPT_F in place of the minimal risk $OPT_F(\mathcal{F})$. Evidently, such a strategy has better convergence property than the previously defined convergent learning strategy, and hence is the expected one in applications.

Accordingly, the specified learning machine and the constructed learning strategy significantly determine the final success of pattern classification. Intrinsically speaking, the learning machine utilized in the VCM can be expressed as follows [6]:

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

where $g(x,\sigma)$ is the Gaussian function

$$g(x,\sigma) = \frac{1}{(\sqrt{2\pi}\sigma)^n} e^{-\|x\|^2/2\sigma^2}. \tag{6}$$

Actually, this learning machine can be formulated by virtue of scale space theory and described by the visual sensation and perception principle: Given a primary image $f(x)$ at a distance of σ from human eyes, the observed blurry image $f(x,\sigma)$ can be mathematically determined by the following partial differential equation([9]):

$$\begin{cases} \frac{\partial f(x,\sigma,)}{\partial \sigma} = \nabla_x f(x,\sigma) \\ f(x,0) = f(x) \end{cases}.$$

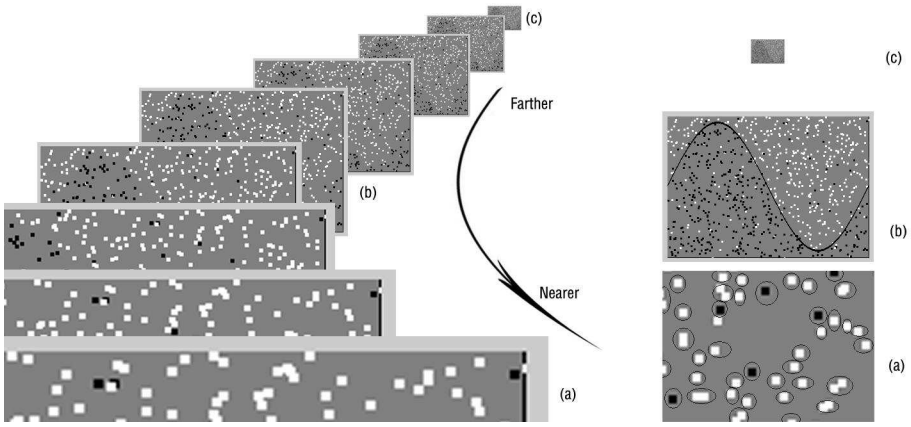

Fig. 1. (a) Observing the data set very closely, a discriminant function composed by the disconnected circles surrounding each datum is perceived; (b) Observing the data from a proper distance, a perfect discriminant function that optimally compromises approximation and generalization performance is perceived; (c) Observing the data set from far away, no discriminant function is perceived.

The solution of the above equation is explicitly expressed as

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

where '$*$' denotes the convolution operation. If the training samples are treated as an imaginary image with expression:

$$f(x,\mathcal{D}_l) = \frac{1}{l}\sum_{i=1}^{l} y_i \delta(x - x_i), \tag{7}$$

then the corresponding blurred image $f(x,\sigma,\mathcal{D}_l)$ at scale σ can be specified by

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

The learning machine $\mathcal{F}_{VCM}$ can naturally be obtained. In fact, the classification performance of the discriminant function $f_{\sigma,\mathcal{D}_l}(x)$ under different values of the parameter σ is in high accordance with the visual phenomenon in observing the classification image by varying its observing distance, as illustrated in Fig. 1.

The purpose of the VCM is then to construct the learning strategy to find the optimal discriminant function from the learning machine $\mathcal{F}_{VCM}$. In fact, any cross validation method, such as the k-fold cross-validation method, can be applied [10–12]. In this method, the given data set is partitioned into k subsamples. Of the k subsamples, a single subsample is retained as the testing data for measuring the learning error of the related discriminant function, and the remaining k-1 subsamples are used as training data. The cross-validation

process is then repeated k times, with each of the k subsamples used exactly once as the testing data. The k results from the folds are then averaged to produce a single learning error. Since the visual validity principle implies that the optimal scales should be selected moderately [13, 14], the appropriate candidates for the scale parameter can be preset and the final result can be obtained by minimizing the cross-validation learning error.

The above strategy defines a mapping $\mathcal{L}_{VCM}$ from the data set $\mathcal{D}_l$ into the $\mathcal{F}_{VCM}$. We will show in the next section that the mapping $\mathcal{L}_{VCM}$ so defined actually achieves a strongly convergent learning algorithm by specifically selecting the scale candidates.

Remark 1. Actually, by applying the Parzen Windows method [15] to estimate the densities underlying the positive class data and the negative class data respectively, and by comparing the estimated densities at every input datum, a classification discriminant function can be obtained. When the method adopts the Gaussian window, the obtained discriminant function is very similar to the function from the VCM. In this sense, the proposed classification implementation scheme and further, the learning machine VCM, are not new. However, in this paper, the learning strategy and the learning machine, which are proposed in visual perspective, can make the following theory more understandable and the analysis of the theory more intrinsic. Furthermore, a significant principle for the new learning theory, visual validity principle, is presented based on the visual phenomenon. Thus, introducing the VCM in the visual perspective will make the related theoretical descriptions more reasonable and natural.

3 Theoretical conclusions on the VCM

In this section, we will show that the proposed learning machine $\mathcal{F}_{VCM}$ converges to the Bayesian estimator. Before presenting the main theorems (proofs are omitted due to limitation of space), we first distinguish two notations of probability spaces.

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

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

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

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

then $\mathcal{P}_2 = (\Omega_X, \mathbb{F}_X, P_X)$ defines another probability space. These two different probability spaces will be used without further justification.

For the given classification problem and 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 is theoretically estimated. The related conclusion is stated as follows:

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 the training sample set generated from P, $E_y(x)$ the average of y at x, and $p(x)$ the density function of x. Assume 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 $\sigma > 0$ and any $\delta \in (0,1)$, $\varepsilon > 0$, there exist positive constants c_1, c_2 c_3, c_4, independent of l and σ, such that*

$$P\{|R(f_{\sigma,\mathcal{D}_l}) - OPT_F| < \varepsilon + P_X\{0 < |E_y(x)p(x)| < Bound(\varepsilon,\delta,l,\sigma)\}\} > 1-\delta \tag{9}$$

where $Bound(\varepsilon,\delta,l,\sigma)$ is defined by

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

Under the mild condition, Theorem 1 implies the following fundamental upper bound estimation on the deviation of the learning function risk from the Bayesian risk in probability

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

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

From (11), in order to maximize the generalization capability of the $f_{\sigma,\mathcal{D}_l}$, or equivalently, to minimize the risk of a discriminant function from the VCM, one possible way is to minimize the function $Bound(\varepsilon,\delta,l,\sigma)$. This provides a strategy of controlling the generalization capability of the learning machine $\mathcal{F}_{VCM}$.

Since ε and δ in (9) are arbitrary, the function $Bound(\varepsilon,\delta,l,\sigma)$ can be regarded as varying with l and σ only. So, with this understanding, some useful observations can be made:

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

These observations show that whenever σ gets to be too large or too small, the function $f_{\sigma,\mathcal{D}_l}$ will never attain good generalization performance. That is, a good (particularly, an optimal) performance can only occur when the scale parameter σ is set as a moderate value. This conclusion is clearly supported by our visual system (the visual validity principle) and coincides completely with the simulations conducted in [6].

A question naturally arisen is: "Is there an optimal scale σ^* and where is it?". The following theorem provides an answer to this question:

Theorem 2. *For any fixed l, the function $Bound(\varepsilon, \delta, l, \sigma)$ has a unique minimum and it attains its minimum 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 appeared in Theorem 1.

Theorem 2 implies that the preferred scale candidates in the VCM should be set with the rank $O(l^{-\frac{1}{2n+2}})$. That is to say, we can set the candidate range of the scale as follows: first, specify the appropriate parameters a and b (say, $a = 10^{-1}$ and $b = 10^3$), and then let $\varepsilon_l = al^{-\frac{1}{2n+2}}$ and $E_l = bl^{-\frac{1}{2n+2}}$, and select the optimal scales in the interval $[\varepsilon_l, E_l]$.

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

$$L_{VCM}(\mathcal{D}_l) = f_{\sigma^*, \mathcal{D}_l}(x) \tag{14}$$

where σ^* is the optimal scale obtained from the following optimization problem:

$$\sigma^* = \arg \min_{\sigma \in [\varepsilon_l, E_l]} CV(\sigma) \tag{15}$$

where $CV(\sigma)$ is the related cross validation error under the scale parameter σ. To minimize the continuous function $CV(\sigma)$ and to find the optimal scale parameter σ^* of (15), any one-dimension global optimization method, such as grid algorithms [16], simulated annealing [17] and evolutionary methods [18], can be adopted.

The following theorem shows that by setting the scale parameter to within $[\varepsilon_l, E_l]$, $\mathcal{L}_{VCM}$ is a strongly convergent learning algorithm for the classification problem.

Theorem 3. *In the setting of Theorem 1, if the optimal scale is selected with rank $O(-\frac{1}{2n+2})$, then for any $\varepsilon > 0$ and $\delta \in (0, 1)$, there is an integer $l(\varepsilon, \delta)$ such that whenever $l > l(\varepsilon, \delta)$*

$$P\left\{R\left(L_{VCM}(\mathcal{D}_l) - OPT_F\right) \geq \varepsilon\right\} < \delta. \tag{16}$$

That is, $\mathcal{L}_{VCM}$ is a strongly convergent learning algorithm for the classification problem.

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

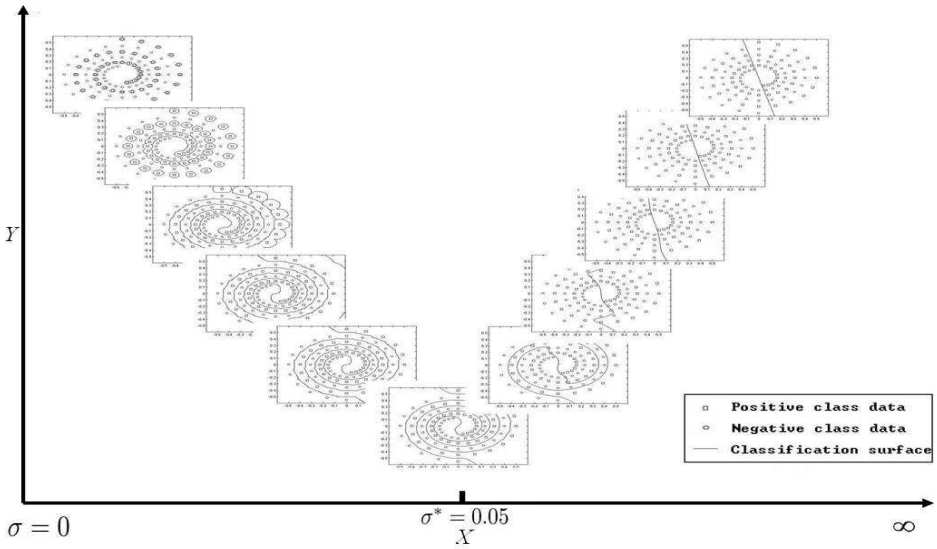

Fig. 2. The generalization capability (Y-axis) of the discriminant function $f_{\sigma,\mathcal{D}_l}$ with the scale parameter σ (X-axis) varying from very small to very large. σ^* is the optimal scale parameter attained by VCM.

4 Simulations and applications in disease diagnosis of the VCM

In this section we provide two sets of numerical simulations to support the feasibility and validity of the theoretical results above, and four applications of the VCM in disease diagnosis. In the experiments, the optimization problem (15) of the VCM was solved by the grid optimization method. All programs were run on the Matlab 7.0 platform in a personal computer with Pentium IV 1.7 CPU, 1 G memory and Windows XP operating system.

4.1 Simulation results

The first set of simulations was conducted to demonstrate the rationality of using $Bound(\varepsilon, \delta, l, \sigma)$ in Theorem 1 to bound the deviation of the risk of a discriminant function $f_{\sigma,\mathcal{D}_l}(x)$ in $\mathcal{F}_{VCM}$ from the Bayesian risk, that is, applying this estimated bound as a measure of the learning capability of $f_{\sigma,\mathcal{D}_l}(x)$. The simulations were carried out by comparing the behavior of $Bound(\varepsilon, \delta, l, \sigma)$ and the performance of the discriminant function $f_{\sigma,\mathcal{D}_l}(x)$. The artificial spiral two-label classification data set $D_{100} = \{x_i^+, +1\}_{i=1}^{50} \cup \{x_i^-, -1\}_{i=1}^{50}$ was used for the comparison, 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 $f_{\sigma,\mathcal{D}_l}$ for each σ is demonstrated in Fig. 2. It can be observed that the generalization capability of $f_{\sigma,\mathcal{D}_l}$ becomes very poor when σ

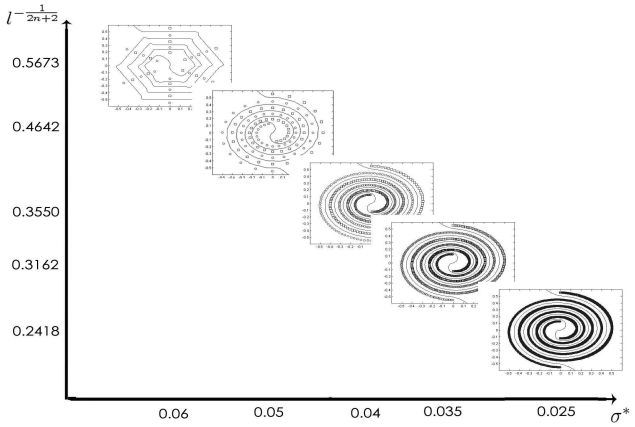

Fig. 3. Performance of the optimal discriminant function $f_{\sigma_l^*,\mathcal{D}_l}$ attained by VCA when data size varies from 30, 100, 500, 1000 to 5000. It should be observed that all $(\sigma_l^*, l^{-\frac{1}{2n+2}})$ are nearly on a line, that is, $l^{-\frac{1}{2n+2}}/\sigma_l^*$ is approximately a constant for any l.

tends to be too large or too small. In fact, for any fixed ε, δ and l, it is easy to deduce that the function $Bound(\varepsilon, \delta, l, \sigma)$ is positive and convex when $\sigma \geq 0$, and its value varies from infinitely large to its finite minimum, and then to infinitely large again, as σ goes from 0 to infinity. The observed performance of $f_{\sigma,\mathcal{D}_l}(x)$ is clearly in accordance with the behavior of $Bound(\varepsilon, \delta, l, \sigma)$ and this verifies the rationality of the developed upper bound estimation (Theorem 1).

The second set of numerical simulations was carried out to verify the correctness 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

$$C = l^{-\frac{1}{2n+2}}/\sigma_l^*. \tag{17}$$

The following spiral data sets $D_l = \{x_i^+, +1\}_{i=1}^{l/2} \cup \{x_i^-, -1\}_{i=1}^{l/2}$ with variable size l was used to test (17), 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, we found the optimal scales σ_l^* to be 0.06, 0.05, 0.04, 0.035 and 0.025 when data size l =30, 100, 500, 1000 and 5000, respectively, as shown in Fig. 3. It is calculated that for each case, $C_{30} \approx 0.1057$, $C_{100} \approx 0.1077$, $C_{500} \approx 0.1126$, $C_{1000} \approx 0.1107$ and $C_{5000} \approx 0.1134$. All are approximately equal. This supports the validity of Theorem 2.

The validity of Theorem 3 will be verified by the application results on disease diagnosis in the next section.

4.2 Disease diagnosis

To further verify the aforementioned theoretical results on the VCM (especially its strong convergence property) and show its performance in bioinformatics

applications, we further applied the VCM to some disease classification problems. In particular, four sets of disease diagnosis data were adopted. The information about the data is listed as follows:

- **Breast cancer data**. This breast cancer 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 (donated by Peter Turney). There are 8 input attributes: number of times being pregnant, plasma glucose concentration a 2 hours in an oral glucose tolerance test, diastolic blood pressure, triceps skin fold thickness, 2-Hour serum insulin, body mass index, diabetes pedigree function, and age. in the binary output class, 0 and 1 mean the non-occurrence and occurrence of the disease, respectively.
- **Heart disease data**. This database contains 76 attributes, but all published experiments only need 14 of them. The 13 input attributes used are respectively age, sex, chest pain type, resting blood pressure, serum cholestoral, 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 variable to be predicted is 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 respectively mean: 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. In the output class, 1, 2, and 3 represents respectively the euthyroidism, hypothyroidism, and hyperthyroidism diagnosis result.

All these data can be downloaded from the UC Irvine Machine Learning Repository (http://www.ics.uci.edu/∼mlearn/MLSummary.html).

To reasonably measure the classification capability of a learning strategy, the following way is generally utilized: first, multiple partitions of the training and testing sets are randomly generated from the original data sets. Second, on each partition, a classifier is trained by the learning strategy and its test set error is obtained correspondingly, the mean and variance of these obtained errors are then taken as the final learning error of the strategy. Based on such ideas, we had conducted three series of simulations (denoted henceforth by A1, A2 and A3 respectively). In series A1, each problem includes 100 partitions of the training and testing sets, which can be directly downloaded from the website: http://ida.first.fhg.de/projects/bench. In series A2 and A3, there are 20 and 10

Table 1. 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 ($A1, A2, A3$)
Broast-cancer	9	200×100,1000×20,2000×10	77×100,385×20,770×10
Diabotis	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 2. Performance comparison between VCM & SVC

Problems	Misclassification rate($A1$)		Misclassification rate($A2$)		Misclassification rate($A3$)	
	SVC(%)	VCM(%)	SVC(%)	VCM(%)	SVC(%)	VCM(%)
Breast-cancer	**25.48±4.41**	25.69±3.38	2.89±0.62	**2.84±0.51**	2.84±0.34	**2.84±0.30**
Diabotis	**23.51±1.48**	25.84±1.61	**0.48±0.54**	0.53±0.56	0.07±0.21	**0.07±0.12**
Heart	**15.62±3.26**	17.19±3.00	**0.30±0.57**	0.45±0.60	**0.0±0.0**	**0.0±0.0**
Thyroid	5.07±2.33	**4.28±1.87**	**0.07±0.30**	0.27±0.82	**0.0±0.0**	**0.0±0.0**
Average	**17.42± 2.87**	18.25±2.47	**0.94±0.51**	1.02 ± 0.62	0.73±0.14	**0.73±0.10**

partitions of the training and testing sets, respectively formed by combining each 5 and 10 training and testing sets of those used in A1. All the data involved are listed in Table 2. The VCM and SVC were respectively applied to these training and testing sets of the four disease diagnosis problems. Specifically, in the implementation of the SVC, the 5-fold cross-validation method was used to select model parameters σ and C from $100 = 10 \times 10$ candidates (10 candidates of C and 10 candidates of σ at each fixed C). The experimental results are summarized in Table 2.

From Table 2, it is easy to observe that when the size of the training data increases, the VCM gradually outperforms the SVC. Particularly, For both series $A1$ and $A2$, the SVC performs better in 3 diagnosis cases, and only in 1 case the VCM performs better. Yet it can also be observed that as the size of the training data becomes larger from $A1$ to $A2$, the advantage of SVM compared with VCM is much smaller. For the $A3$ case, where the size is largest, the VCM evidently outperforms the SVC. In particular, in all of 4 applications, the misclassification rates of the VCM is no larger than those of the SVC. This shows the success of the VCM in disease diagnosis, especially in view of its strong convergence property.

5 Conclusions

In this paper, we have introduced a visual classification method and shown its strong convergence. It has been supported by several simulations and real-life applications in disease diagnosis. In particular, the theoretical conclusion on its achieving the posterior expected minimal value, i.e., the Beyesian estimator, has been illustrated. This theoretical result has been verified by a series of synthetic

simulations and applications in disease classification. Compared with the well-known support vector classification method, the VCM has been shown to be effective and efficient. The proposed method thus provides useful techniques in the analysis of bioinformatic data.

References

1. L. Holmstrom, P. Koistinen, J. Laaksonen, and E. Oja. Neural and statistical classifiers - Taxonomy and two case studies. *IEEE Transactions on Neural Networks*, 8(1): 5-17, 1997.
2. M.H. Yang, D.J. Kriegman and N.Ahuja. Detecting Faces in Images: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intellegence*, 24:34-58, 2002.
3. F. Zhang, X.Z. Qiao, and Z.Y. Liu. Parallel divide and conquer bio-sequence comparison based on Smith-Waterman algorithm. *Science in China Series F*, 47(2):221-231, 2004.
4. L.J. McGuffin, K. Bryson, and D.T. Jones. The PSIPRED protein structure prediction server. *Bioinformatics*, 16(4): 404-405, 2000.
5. M.E.M. Noble, J.A. Endicott, and L.N. Johnson. Protein Kinase Inhibitors: Insights into Drug Design from Structure. *Science*, 303(5665): 1800-1805, 2004.
6. Z.B. Xu, D.Y. Meng and W.F. Jing. A New Approach for Classification: Visual Simulation Point of View. *Lecture Notes in Computer Science 3497*, Part II: 1-7, 2005.
7. Yee Leung, Jiang-She Zhang,and Z.B. Xu. Clustering by scale space filtering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22:2-15,2000.
8. V.N. Vapnik. *Statistical Learning Theory*. J. Willey, New York. 1998.
9. Witkin. Scale-space filtering. *Proc. 8th IJCAI*, 1019-1022, 1983.
10. M. Kearns, and D. Ron. Algorithmic stability and sanity-check bounds for leave-one-out cross-validation. *Neural Computation*, 11 (6): 1427-1453, 1999.
11. Y. Bengio, and Y. Grandvalet. No unbiased estimator of the variance of K-fold cross-validation. *Journal of Machine Learning Research*, 5: 1089-1105, 2004.
12. S. Acton and A. Bovik. Piecewise and local image models for regularized image restoration using cross-validation. *IEEE Transactions on Image Processing*, 8(5):652 - 665, 1999.
13. S.Coren, L.Ward, and J.Enns. *Sensation and Perception*. Harcourt Brace College Publishers,1994.
14. E.B. Goldstein. *Sensation and Perception*. Wadsworth Thomson Learning, 2002.
15. R.O. Duda, P.E. Hart, and D.G. Stork. *Pattern Classification*. Wiley, New York, 2004.
16. C. Padgett, and K. KreutzDelgado. A grid algorithm for autonomous star identification. *IEEE Transaction on Aerospace and Electronic Systems*, 33(1): 202-213, 1997.
17. H. Haario, Esakesman. Simulated annealing process in general state space. *Adv. Appl. Prob.*, 23:866-893, 1991.
18. John H. Holland. *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control and Artificial Intelligence*. MIT Press, Cambridge, MA, second edition, 1992.