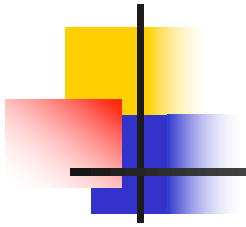
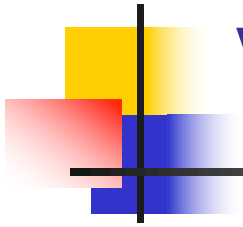
Visual Analytics



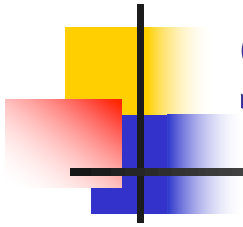
Outline

- Introduction
- The existing approaches
- Visual sensation principle
- Visual classification approach
- Visual learning theory
- Concluding remarks



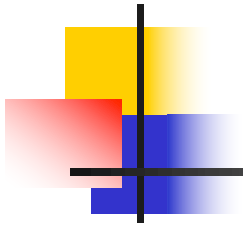
Visual Analytics

- "the science of analytical reasoning facilitated by visual interfaces"
- facilitates analysis and understanding of large, complex and ill-defined data across multiple domains
- uses information technology to analyze and interpret enormous amounts of disparate conflicting, incomplete, often unreliable, and constantly changing data to present it visually for evaluation
- processes & goals of analysis dominate



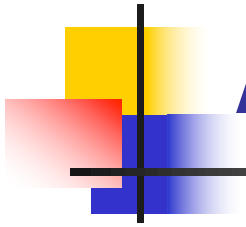
Scenario

- Consider the intelligence community
  - hundreds of thousands of documents
    - many types of documents
    - many languages
  - 6 month timeframe
- Processing
  - graphical display of document clusters by
    - region
    - date
    - severity
  - trying to detect patterns



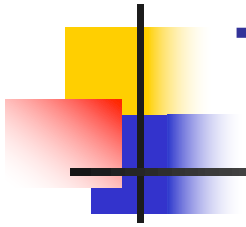
Data Properties

- wide ranging & huge volume
  - audio, video, imagery
  - reports - institutional and government
  - newspapers
  - sensors
  - digital & analog
  - multidimensional, multi-source, time varying
  - e-mail, blogs, social networks, ...
- disparate, conflicting, and dynamic
- incomplete, inconsistent, and deceptive



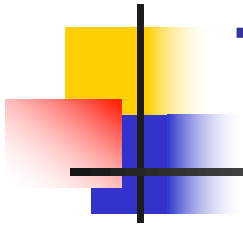
Analyst

- Forage complex and contradictory bodies of information
- Seek detailed knowledge of specific facts to
  - support or refute candidate hypothesis
  - identify and bridge knowledge gaps
  - discover previously unknown relationships and evidence
- Under stressful, time-limited, uncertain conditions with variability in completeness and accuracy of data



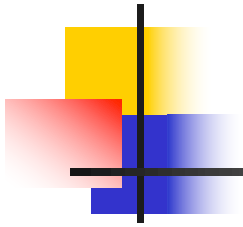
Technologies

- Knowledge Management
  - managing the learning process
- Statistical Analysis
  - average, distribution, correlation
- Cognitive Science
  - analysis of cognitive dissonance and effort
- Decision Science
  - managing the processing of deciding how to explore the data space



Tasks

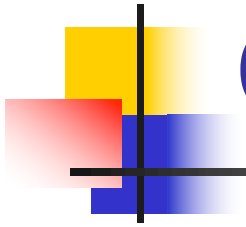
- Analysis and Discovery
- Synthesis to Derive Insight
- Detect Unexpected Event/Results
- Communication of Analytical Results



Focus

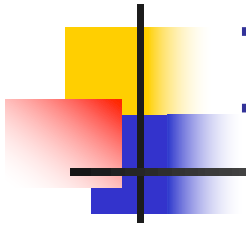
- Analytical Reasoning
- Visual Representation and Interaction
- Data Representation and Transformation
- Result Production, Presentation, & Dissemination

Analytical Reasoning: Challenge

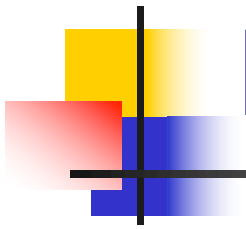


- Applying human judgments to reach conclusions from a combination of evidence and assumptions.
- VA facilitates AR through creation of software that maximizes human capacity to perceive, understand, and reason about complex and dynamic data and situations.
- Thinking
  - Convergent
  - Divergent

Visualization Representation & Interaction Challenges

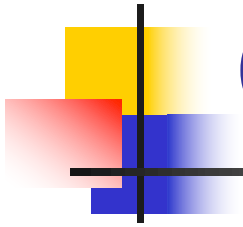


- Representations
  - based on scientific principles for depicting information
- Interaction
  - must take into account the time constraints associated with varying levels of urgency in an analytic task.



Data Challenges

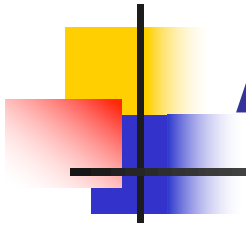
- Vast Streams of all Types
  - large, complex, and dynamic collection
- Representations & Transformations
  - convert all types of conflicting and dynamic data into forms that facilitate analytical understanding
  - support varying levels of abstraction to support scale and context (or lack thereof)



Communication

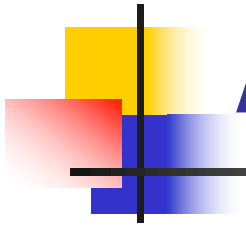
- Production
  - analytical assessments and output assembled into presentations
- Presentation
  - relevant facts and evidence, results, and uncertainties
- Dissemination
  - numerous audiences
    - policy makers to general population

Query-driven Visualization & Analytics

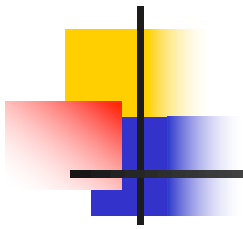


- Capabilities for knowledge discovery and hypothesis testing to understand phenomena “hidden” in vast and complex collections of scientific data.
- Instead of making a single image of a terabyte’s worth of data, we compute an “image” that contains only data deemed to be “scientifically interesting”

Query-driven Visualization & Analytics cont.

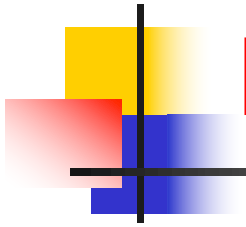


- Reduces computation, cognitive, & visual load
- Major new capability is having to find and process only the small subset of data that is “interesting”
- Combines technologies from scientific visualization, visual analytics, and scientific data management



Emerging Core Technologies

- Multiscale information representation and visualization
  - analysis ranging from molecules to ecosystems
- Analysis of massive unstructured text
  - requires data signatures to develop high dimensional representations for statistical and semantic analysis - relationship discovery
- Temporal analytics
  - time change of structure, relationship & meaning

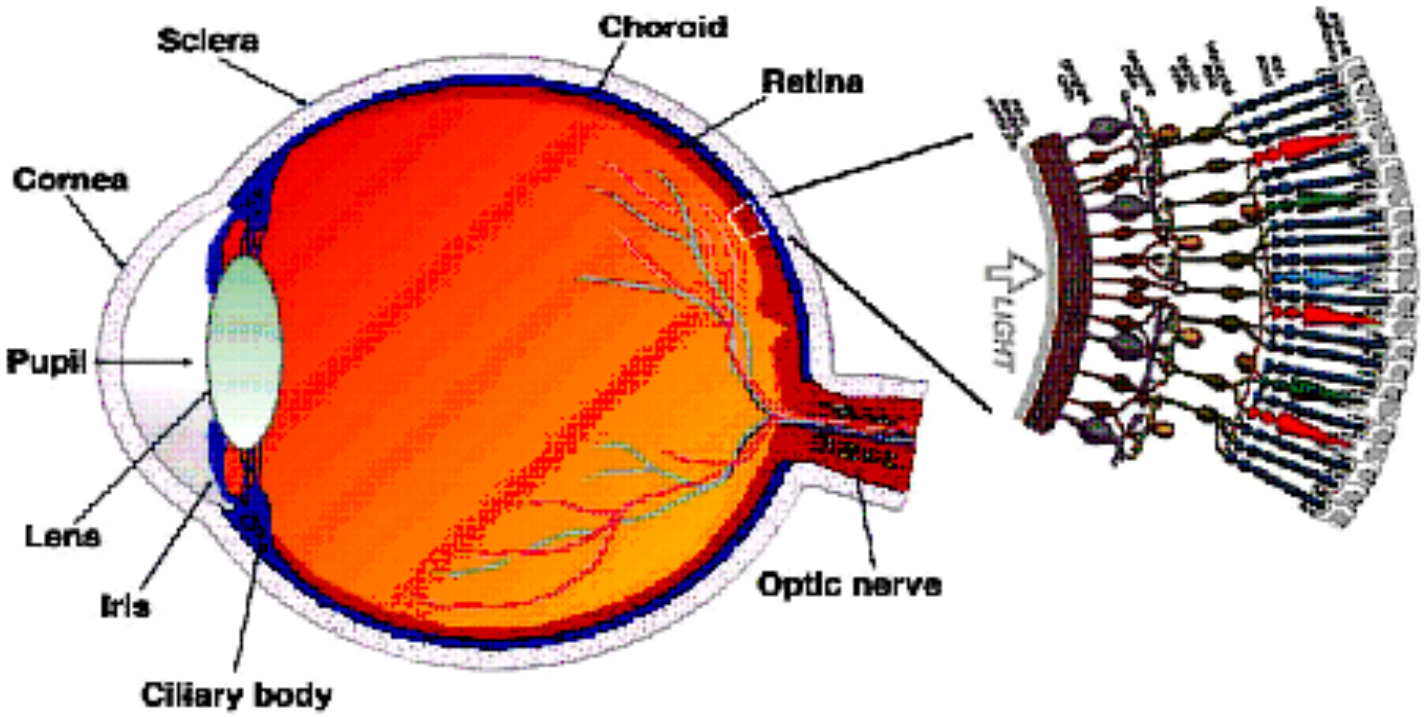


Example

- The essence of DM is modeling from data. It depends not only on how the data are generated, but also on how we sense or perceive the data. The existing DM methods are developed based on the former principle, but less on the latter one.
- Our idea is to develop DM methods based on human visual sensation and perception principle (particularly, to treat a data set as an image, and to mine the knowledge from the data in accordance with the way we observe and perceive the image).
  - IEEE PAMI 2012

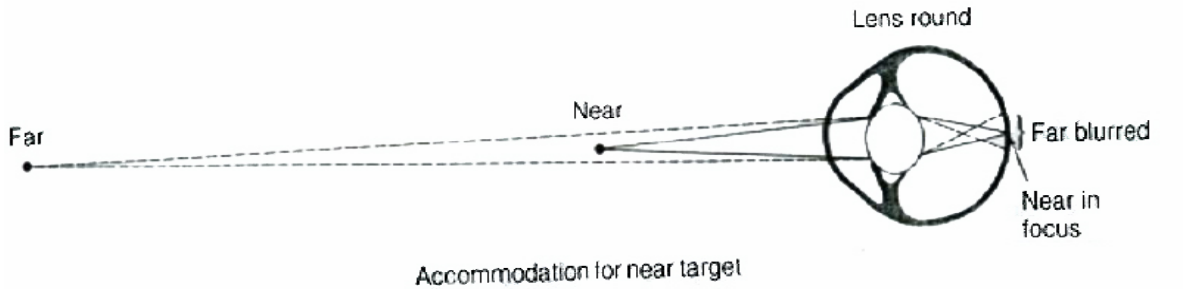
2.1. Visual sensation principle

The structure of the human eye

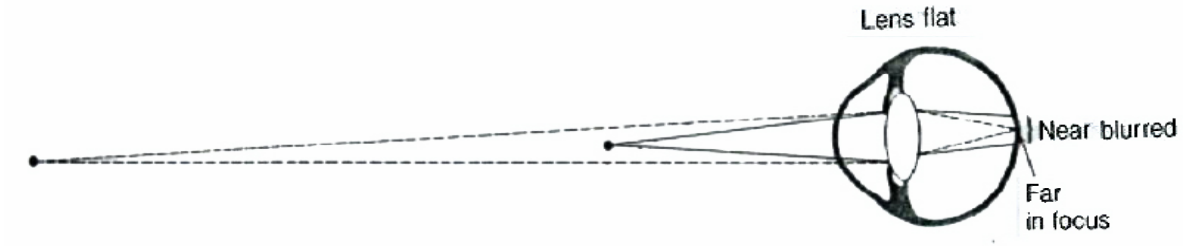


2.1. Visual sensation principle

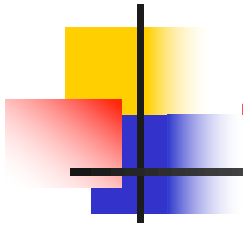
Accommodation (focusing) of an image by changing the shape of the crystalline lens of the eyes (or equivalently, by changing the distance between image and eye when the shape of lens is fixed)



Accommodation for near target



Accommodation for far target



2.1. Visual sensation principle

How an image in retina varies with the distance between object and eye (or equivalently, with the shape of crystalline lens)?

Scale space theory provides us an explanation. The theory is supported by neurophysiologic findings in animals and psychophysics in man directly.

2.2. Scale Space Theory

- Let $p(x): R^n \to R$ denote an image whose gray degree at x is $p(x)$. $P(x,\sigma)$ is the image appeared in the retina at the scale σ. The scale space theory then says that $P(x,\sigma)$ obeys to

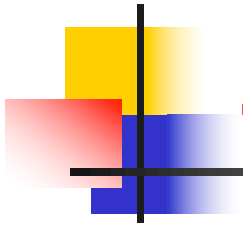
$$\begin{cases} \dfrac{\partial P}{\partial \sigma} = \Delta_x P \\ P(x,0) = p(x) \end{cases}$$

or equivalently,

$$P(x,\sigma) = p(x) * g(x,\sigma) = \int_{\Omega} g(x-y)p(y,\sigma)dy\,,$$

where '*' denotes the convolution operation and $g(x,\sigma)$ is Gaussian kernel defined by

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



2.2. Scale Space Theory

- In image processing terms, this means that the image of an object reflected in the retina at scale σ, $P(x,\sigma)$, is the blurred result of the original image $p(x)$. The blurring process is coincidently the heat diffusion process.
- This is derived under certain isotropy assumption; otherwise more complicated PDE model can be formulated, say, Perona & Malik model.

2.2. Scale Space Theory

- Perona & Malik model (1990)

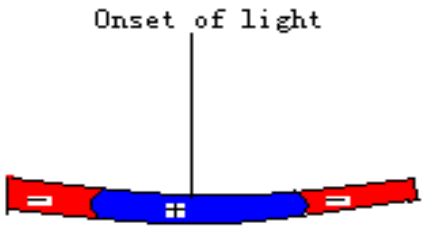
$$\begin{cases} \dfrac{\partial u}{\partial t} = div(c(|\nabla u|^2)\nabla u) & in \quad \Omega \times (0,T) \\ u(0,x) = u_0(x) & in \quad \Omega \\ \dfrac{\partial u}{\partial n} = 0 & on \quad \partial\Omega \times (0,T) \end{cases}$$

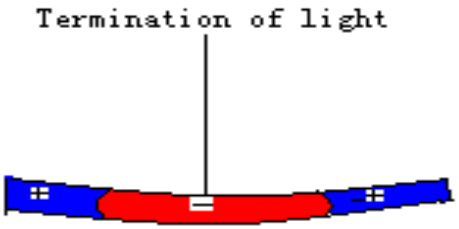
Perona & Malik model is anisotropic. It can give prominence to the edges of the image, but has no explicit expression of its solution.

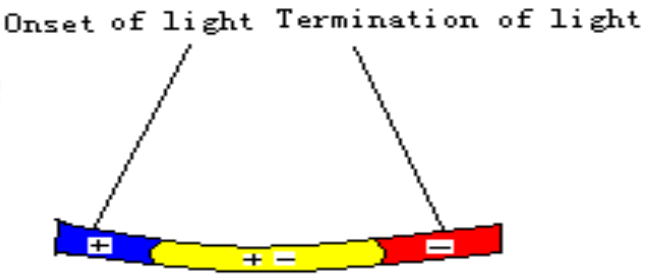
2.3 Cell responses in retina

Only change of light can be perceived and only three types of cell responses exist in retina:

- 'ON' response: the response to arrival of a light stimulus (the blue region)
- 'OFF' response: the response to removal of a light stimulus (the red region)
- 'ON-OFF' response: the response to the hybrids of 'on' and 'off' (because both presentation and removal of the stimulus may simultaneously exist) (the yellow region)

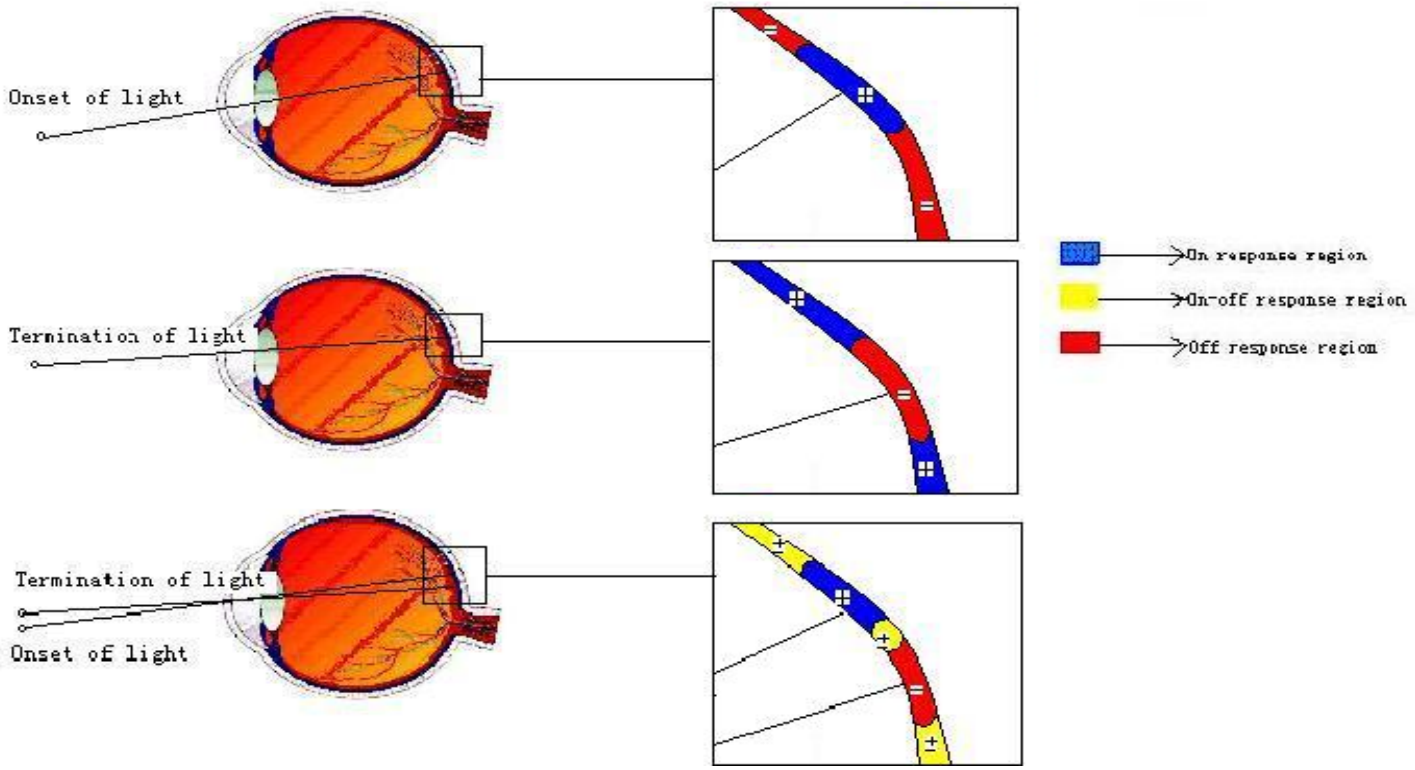






2.3. Cell responses in retina

Between on and off regions, roughly at the boundary is a narrow region where on-off responses occur. Every cell has its own response strength, roughly, the strength is Gaussian-like.



3. Visual Classification Approach: our philosophy

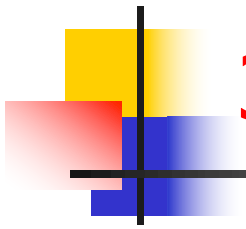
How to bridge visual sensation principle and classification?

- **To view a given data set as an image (imaginary one)**
  See every positive/negative training sample x_i as a spot light with unit strength $\delta(x-x_i)$ / $-\delta(x-x_i)$, causing an "on"/"off" responses in the retina. (where $\delta(x)$ is a dirac function). Consequently all the data forms an image

$$p(x, D_l) = \frac{1}{N^+ + N^-}\left(\sum_{i=1}^{N^+} \delta(x - x_i^+) + \left(-\sum_{i=1}^{N^-} \delta(x - x_i^-)\right)\right)$$

- **To obtain a family of images by using scale space formulation**

$$P(x, \sigma, D_l) = p(x, D_l) * g(x, \sigma)$$



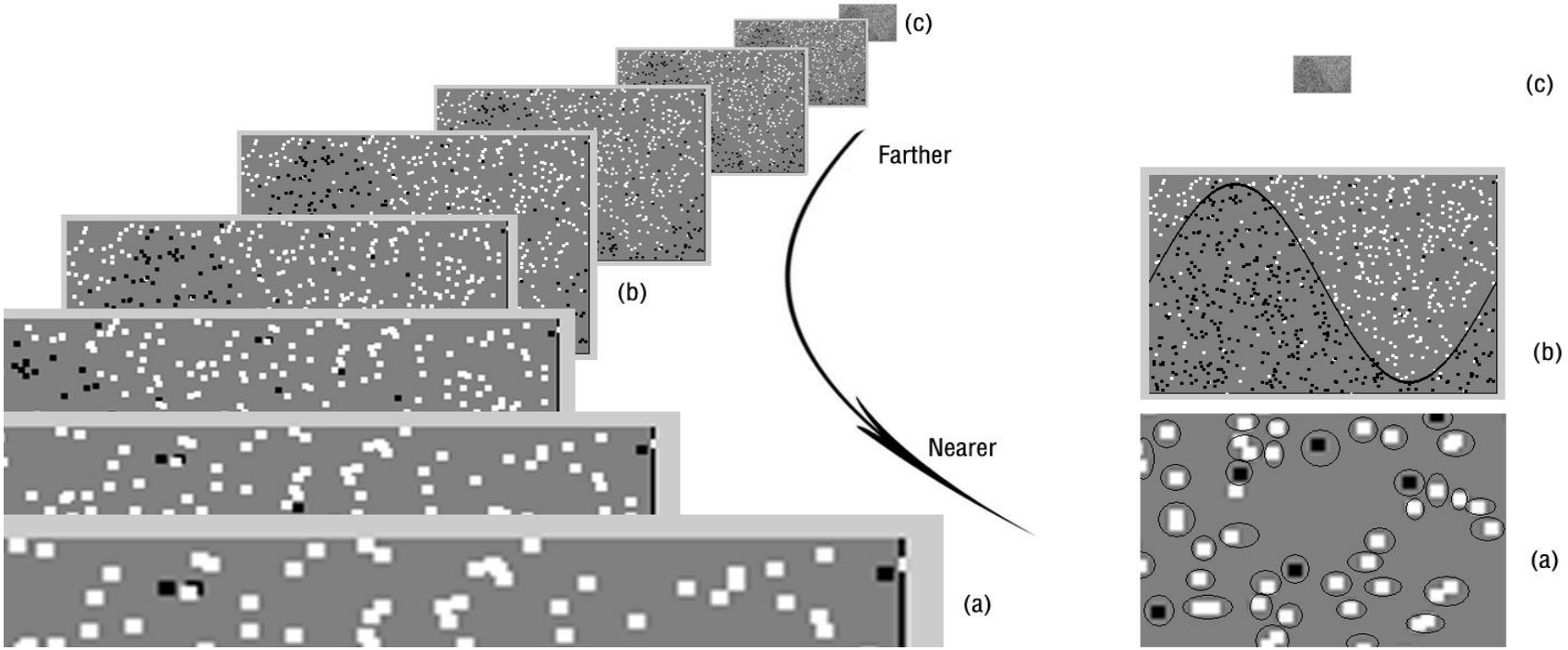
3. VCA:

- For any fixed scale σ_0, view the "+"/"-" class as the "on"/"off" response region, and the boundary as the "on-off" region. Or correspondingly, the discriminant function is defined by $f_{\sigma_0,D_i}(x) = \mathrm{sgn}(P(x,\sigma_0,D_i))$, namely, the classification boundary by $\{x : P(x,\sigma_0,D_i) = 0\}$.

The key is how to choose an appropriate scale σ^*.

3. VCA: A method to choose scale

An observation



3. VCA: A method to choose scale

- There exits an $\varepsilon > 0$ such that when $\sigma < \varepsilon$, $f_{\sigma, D_l}(x)$ can classify every input samples correctly, but not any or has very poor generalization capability;
- $\lim_{\sigma \to \infty}(P(x, \sigma, D_l))$ is a uniform distribution in the input space, so there is a large positive number N such that whenever $\sigma > N$, $P(x, \sigma, D_l)$ has no or very poor approximation capability to the training data.
- This shows that the expected scale is within a bounded interval $[\varepsilon, N]$(too large or too small is meaningfulness).
- There are still infinite possibilities of scales to choose. If only finite number of scales is selected, the well-known cross-validation approach can be applied, to make the search of σ^* feasible.

3. VCA: Method to choose scale

A discretilization scheme for scales:

- Weber's Law provides us a very useful cue:

Weber's law in physiology: a person can not recognize the difference between two images whose fraction for line length of the scale parameters is smaller than 0.029.

This suggests the following discretilization scheme:

$$\Delta\sigma = 0.029$$

- Based on the above discretilization scheme, a finite set of discreminant functions $\{f_{\sigma_i, D_i}(x): i=1,2,\ldots,M\}$ $(M=(N-\varepsilon)/\Delta\sigma)$ is obtained.
- Any cross-validation approach (say, Leave One Out or 5-fold across method) can be applied to guide the search of the optimal scale σ^*.

3. VCA: Procedure

Visual Classification Procedure

Step I (*Initialization*): Choose the parameters ε and N, obtain the imaginary image:

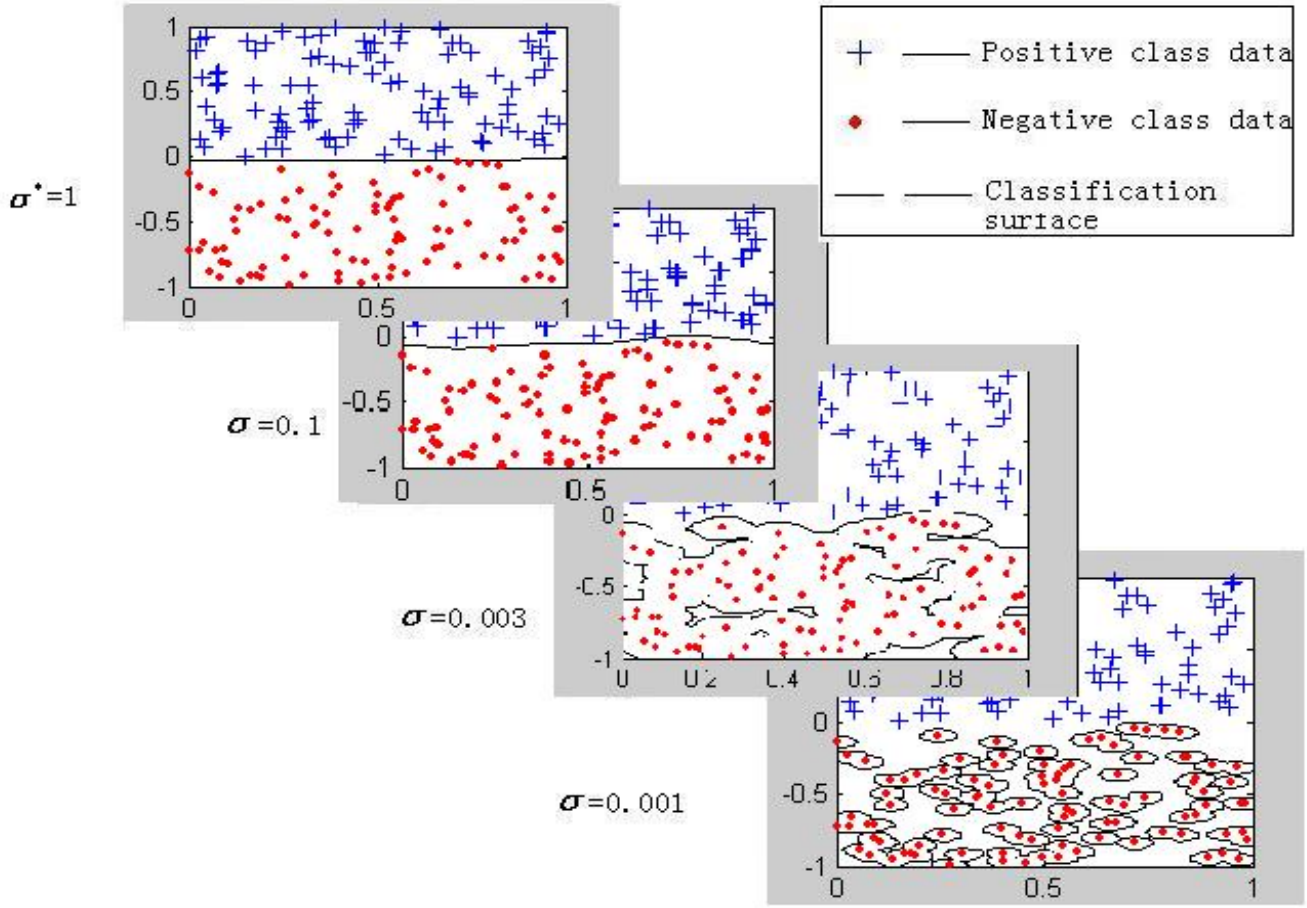
$$p(x, D_l) = \frac{1}{N^+ + N^-}\left(\sum_{i=1}^{N^+} \delta(x - x_i^+) + \left(-\sum_{i=1}^{N^-} \delta(x - x_i^-)\right)\right);$$

Step II (*Scale determination*): Apply a cross-validation method combined with the discretization scheme deduced from Weber's law to determine an appropriate scale σ^* in $[\varepsilon, N]$;

Step III (*Visual classification*): Compute the blurred image $P(x, \sigma^*, D_l) = p(x, D_l) * g(x, \sigma^*)$, and then find the discriminant function $f_{\sigma^*, D_l}(x) = \mathrm{sgn}(P(x, \sigma^*, D_l))$.

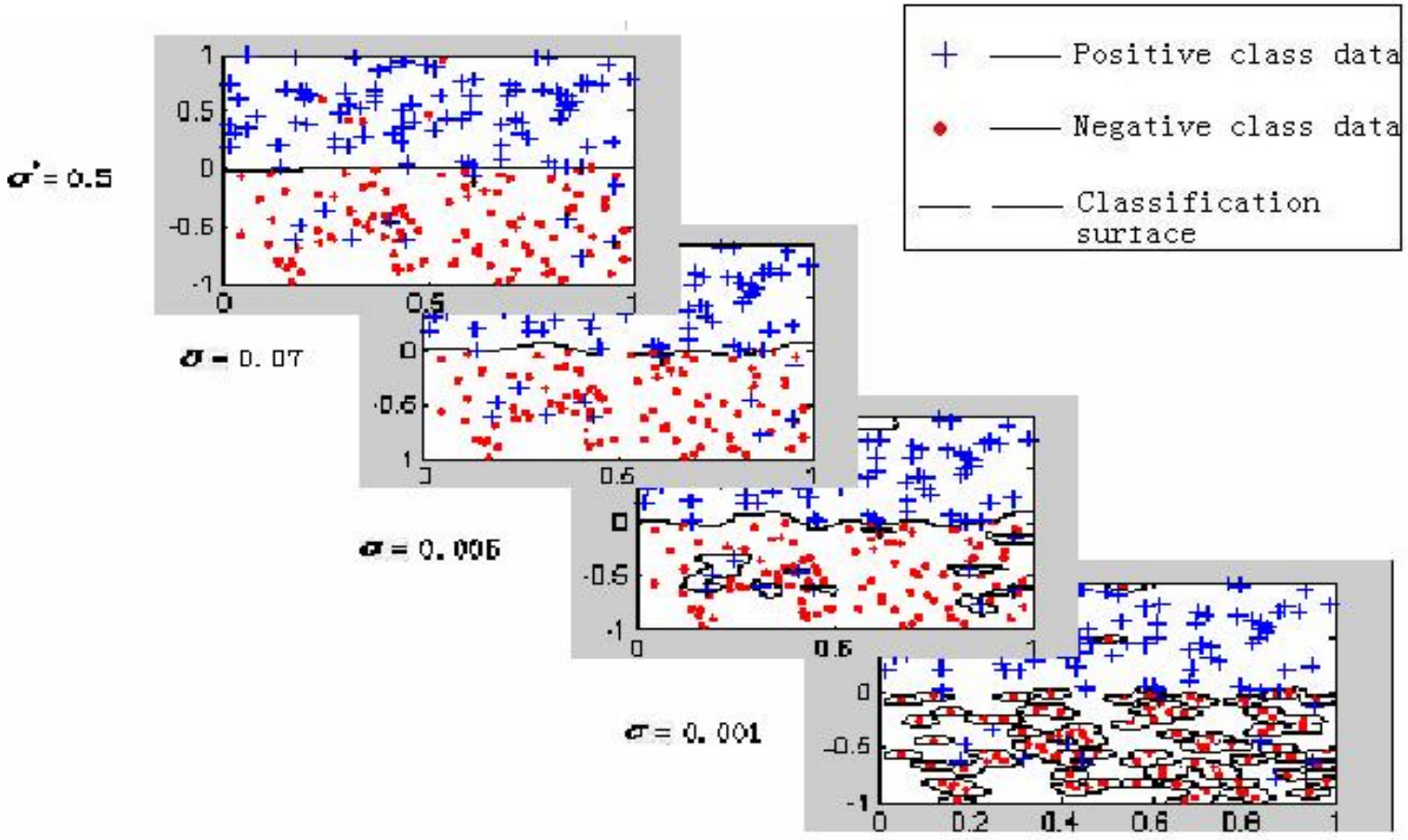
3. VCA: Demonstrations

Linearly separable data without noise



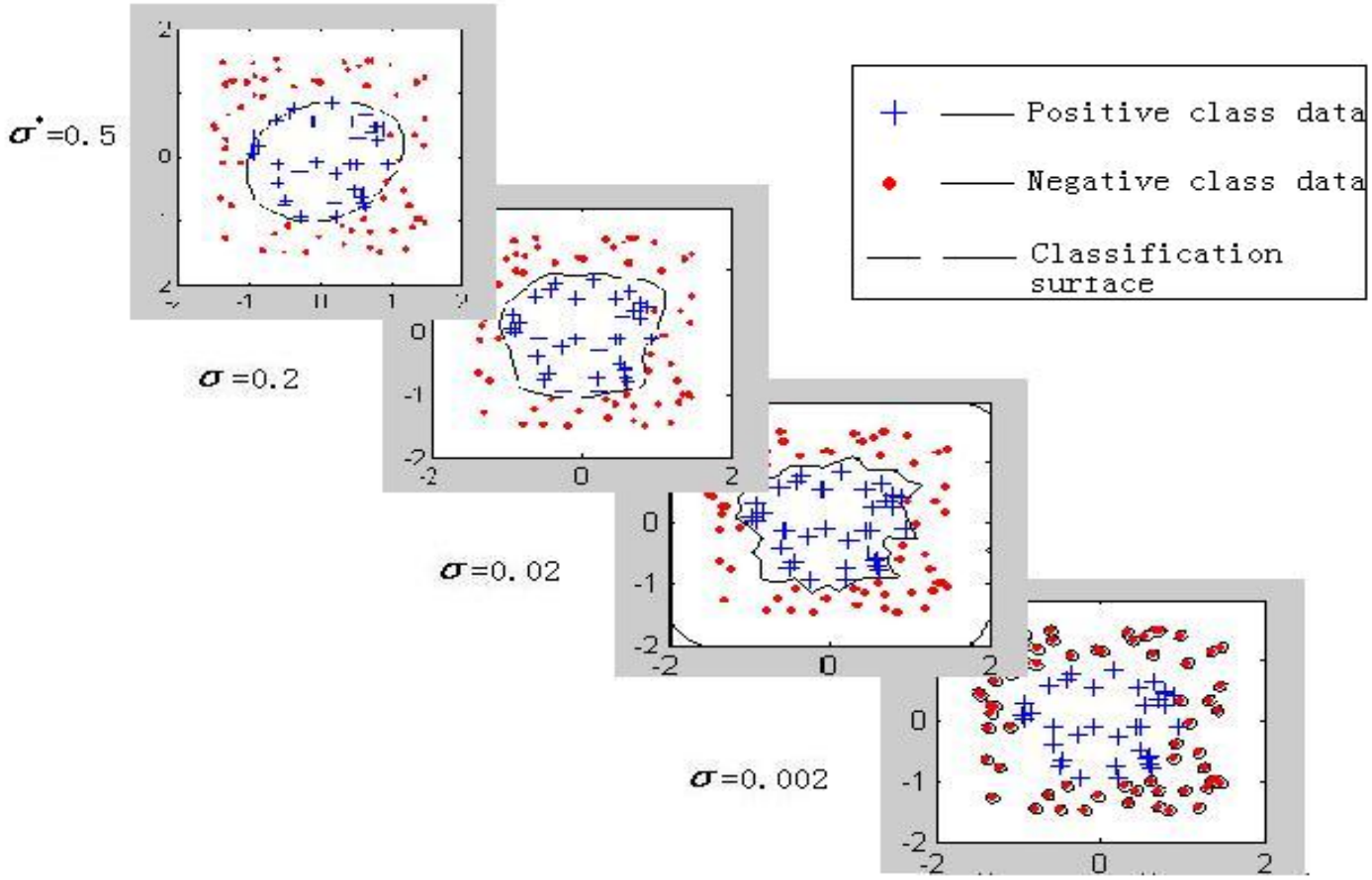
3. VCA: Demonstrations

Linearly separable data with 5% noise



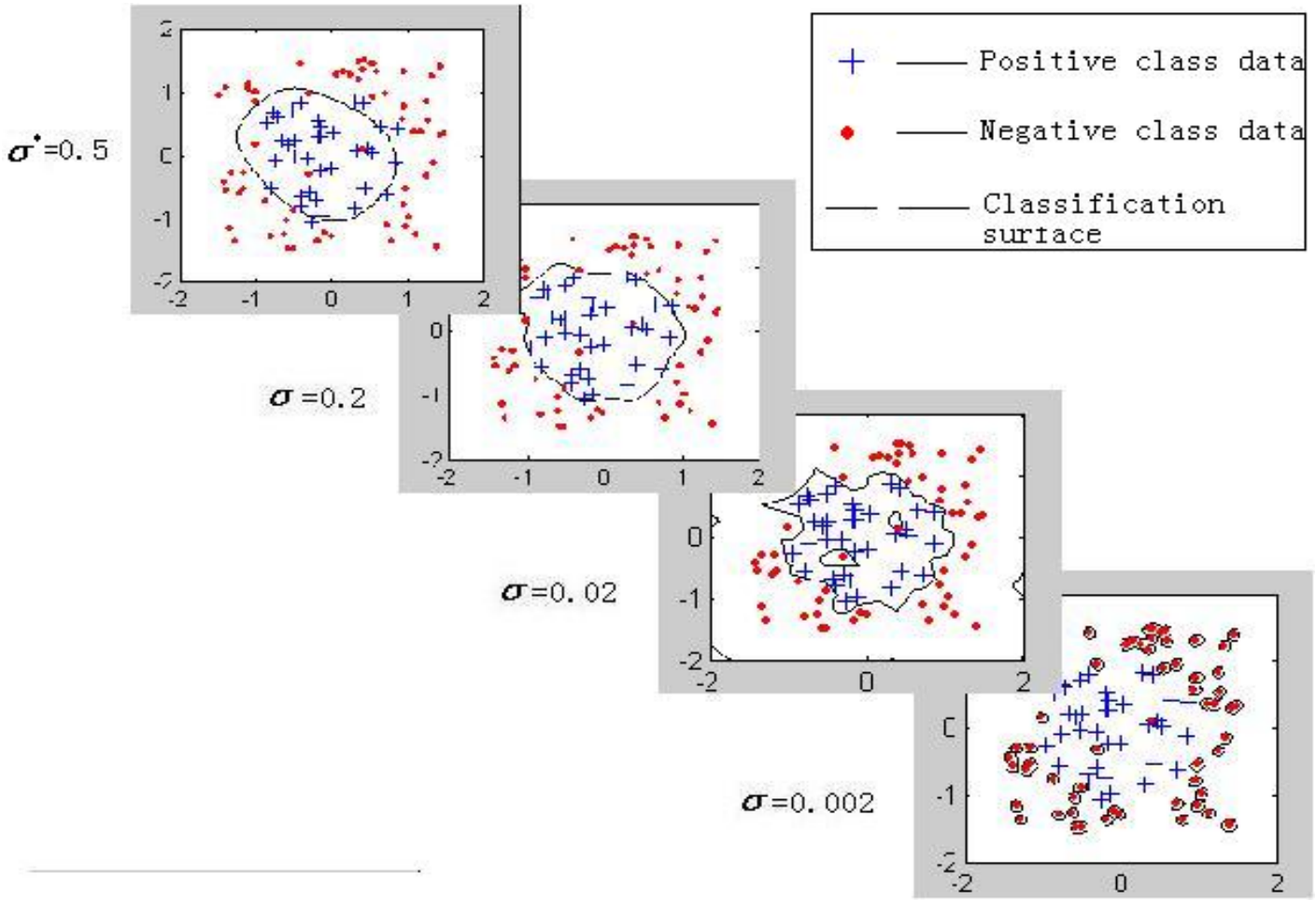
3. VCA: Demonstrations

Circularly separable data without noise



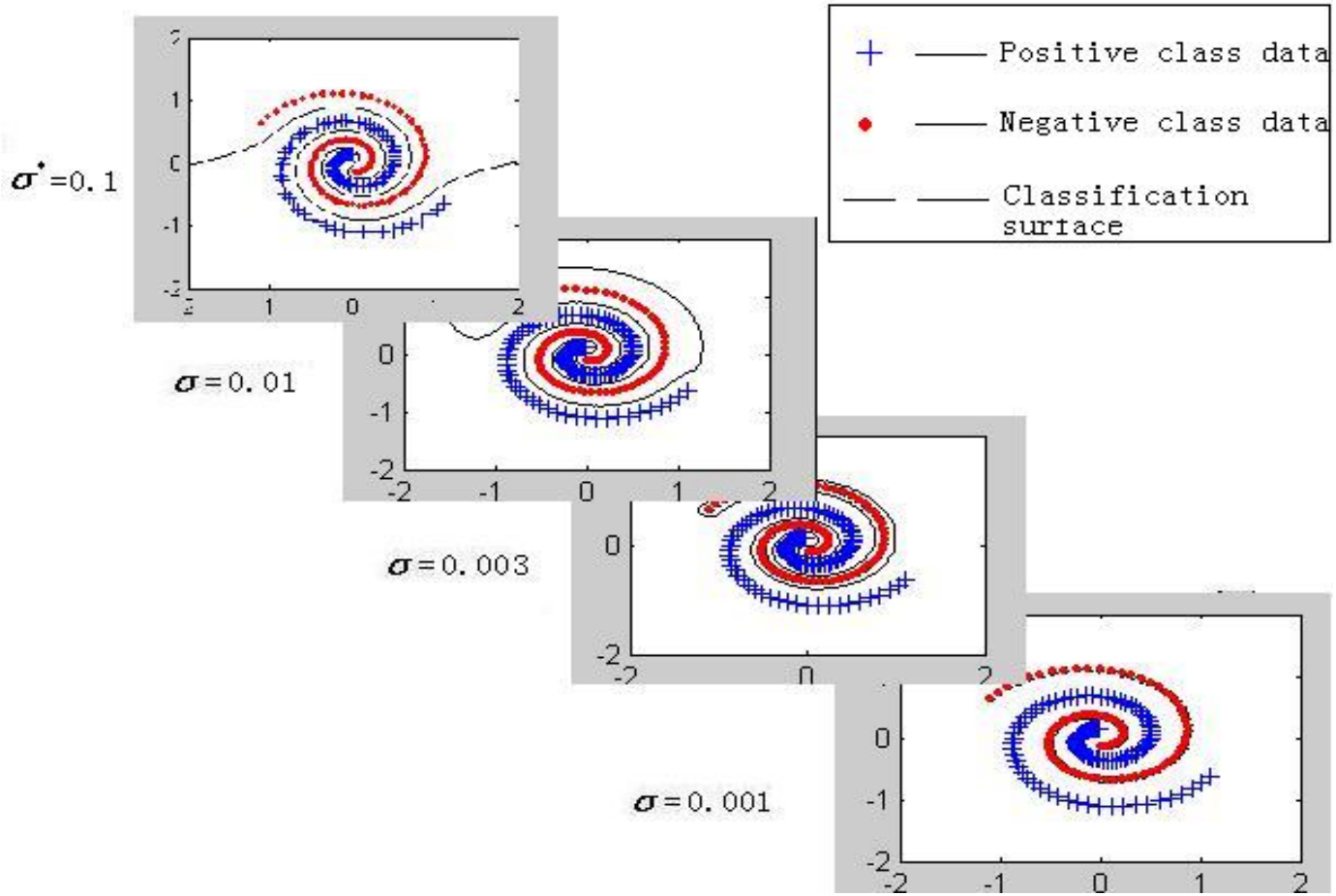
3. VCA: Demonstrations

Circularly separable data with 5% noise



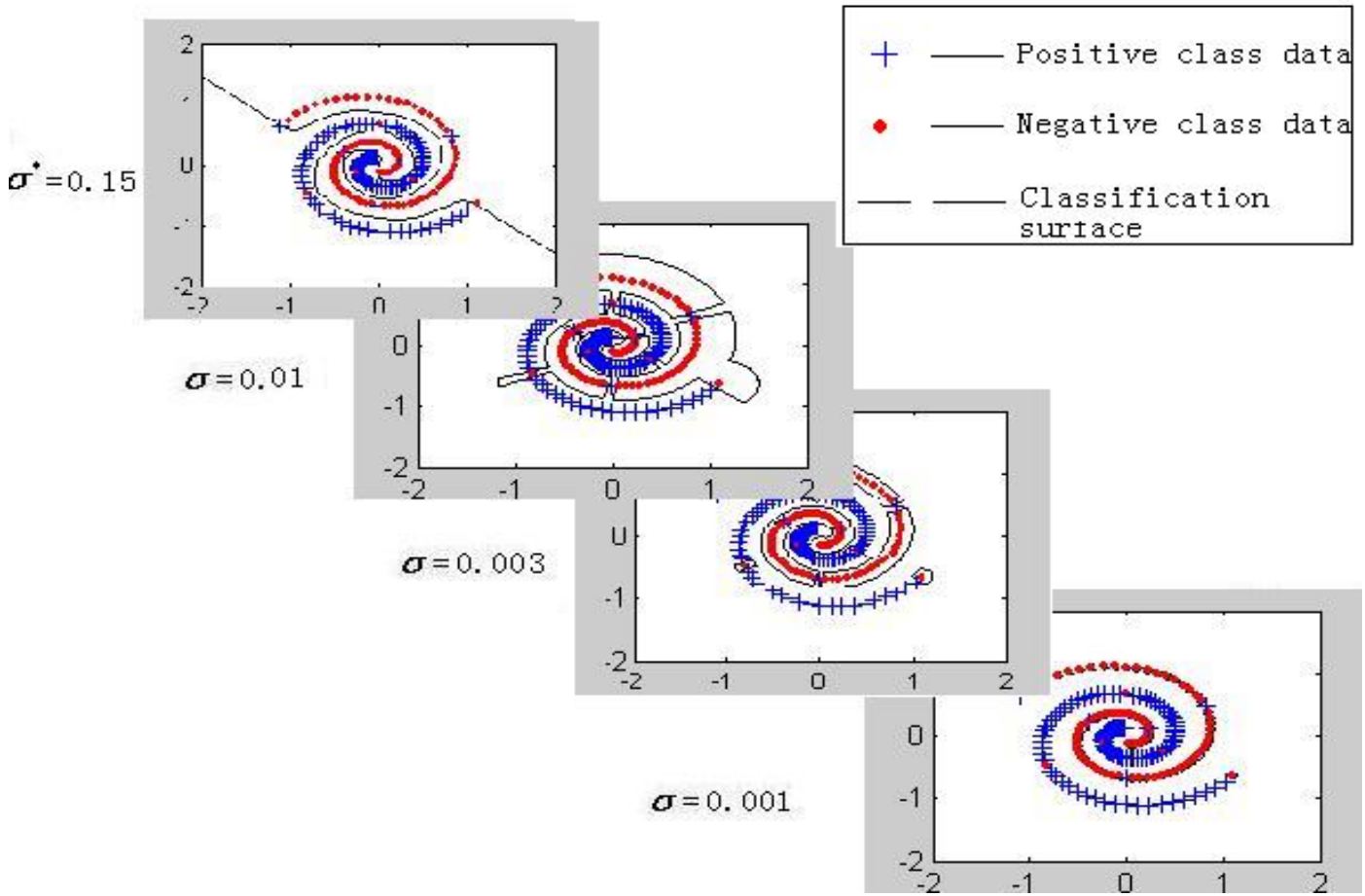
3. VCA: Demonstrations

Spirally separable data without noise



3. VCA: Demonstrations

spirally separable data with 5% noise



3. VCA: Efficiency test

11 groups of benchmark datasets from UCI, DELVE and STATLOG

	Input Dim	Size of training set	Size of test set
banana	2	400	4900
Breast-cancer	9	200	77
diabetis	8	468	300
Flare-sola	9	666	400
german	20	700	300
heart	13	170	100
image	18	1300	1010
thyroid	5	140	75
titanic	3	150	2051
twonorm	20	400	7000
waveform	21	400	4600

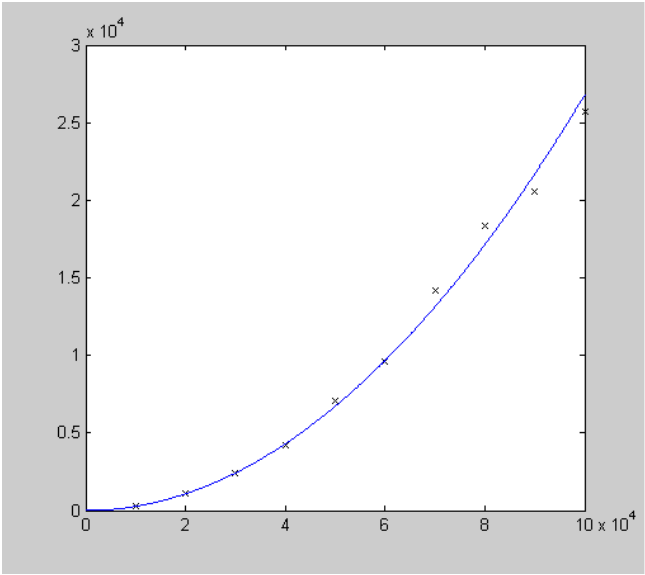
3. VCA: Efficiency test

Performance comparison between VCA & SVM

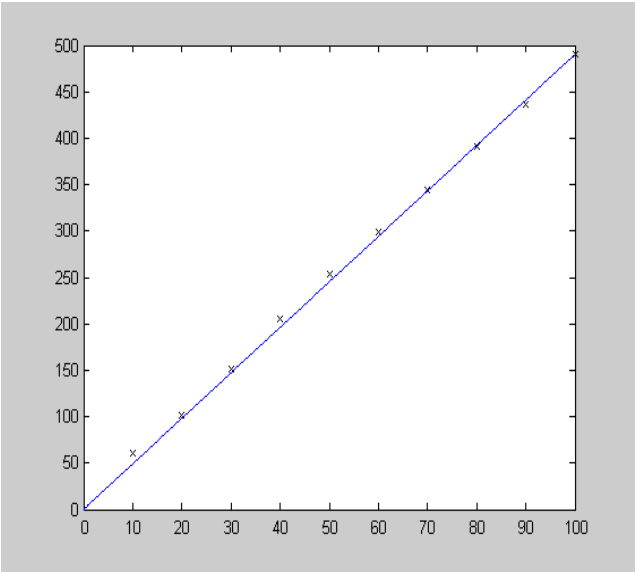
Problems	$t = t_{ps} + t_{tr}$		Misclassification rate	
	SVM(s)	VCA(s)	SVM(%)	VCA(%)
Banana	37579.2=37512.5+66.7	375.10	10.54±0.49	10.81±0.51
Breast-cancer	6446.38=6445+1.38	314.60	25.48±4.41	24.82±4.07
Diabetis	65270.24=65257.5+12.74	665.50	23.51±1.48	25.84±1.81
Flare-solar	170190.5=170165+25.53	1173.70	32.37±1.80	35.01±1.72
German	203328.3=203307.5+20.81	1984.40	23.59±2.19	25.27±2.39
Heart	4486.83=4485+1.83	205.70	15.62±3.26	17.22±3.51
Image	1137431=1137300+131.4	6279.90	2.97±0.57	3.62±0.63
Thyroid	3071.06=3070+1.06	157.30	5.07±2.33	4.35±2.34
Titanic	3366.14=3362.5+3.64	169.40	22.92±1.16	22.31±1.00
Twonorm	37152.53=37082.5+70.03	762.30	2.52±0.15	2.67±0.39
Waveform	38282.77=38220+62.77	90.75	10.52±0.46	10.64±0.98
Mean	155145.9=155109.8+36.1	1107.15	15.92±1.64	16.60±1.76

3. VCA: Scalability test

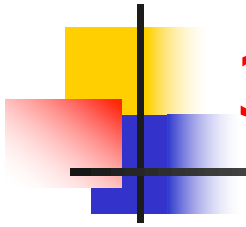
Time complexity of VCA with increase of size of training data is quadratic (a), with increase of dimension of data is linear (b).



(a): fixed 10-D but varying size data sets are used.



(b): Fixed 5000 size but varying dimension datasets are used.



3. VCA: conclusion

1. Without increase of misclassification rate (namely, loss of generalization capability), much less computation effort is paid, as compared with SVM (approximately 0.7% times of SVM is required, increasing 142 times computation efficiency). That is, VCA has very high computation efficiency.

2. The VCA 's training time increases linearly with dimension and quadtratically with size of training data. This shows that VCA has a very good scalabity.

4. Theory: Visual classification machine

- Formalization (Learning theory)

Let $Z = X \times Y$ be sample space (X be pattern space and $Y \in \{-1, 1\}$ label space), and assume that there exists a fixed but unknown relationship *F* (or equivalently, there is a fixed but unknown distribution $\tilde{F}(x, y) = \bar{F}(x)F(y/x)$ on Z).

Given a family of functions

$$H = \{f(x, \alpha),\ \alpha \in \Lambda\}$$

and a finite number of samples

$$D_l = \{x_i, y_i\}_{i=1}^{l} \subset Z$$

which is drawn independently identically according to $\tilde{F}$.

4. Theory: Visual classification machine

- Formalization (cont.)

We are asked to find a function $f^* = f(x, \sigma^*)$ in H which approximates F in Z, that is, find a a function f^* in H, for a certain type of measure Q between machine's output $f(x, \alpha)$ and actual output y, so that

$$f^* = \arg\inf_{f \in H}(R(f)) \quad \text{(Learning problem)}$$

where

$$R(f(x,\alpha)) = \int_Z Q(y, f(x,\alpha)) d\tilde{F}(x,y).$$

(risk or generalization error)

4. Theory: Visual classification machine

- Learning algorithm (Convergence)

A learning algorithm *L* is a mapping from D_l to *H* with the following property:

For any $\varepsilon > 0, \delta \in (0,1)$ *, there is an integer* $l(\varepsilon, \delta)$ *such that whenever* $l > l(\varepsilon, \delta)$ *,*

$$P\{\left|R(L(D_l)) - OPT_F(H)\right| < \varepsilon\} \geq 1 - \delta$$

where $OPT_F(H) = \inf_{f \in H} R(f)$.

In this case, we say that *L(Z)* is a (ε, δ)-solution of the learning problem. Given an implementation scheme of a learning problem, we say it is convergent if it is a learning algorithm.

4. Theory: Visual classification machine

- Visual classification machine (VCM)
  - The function set

$$H = \{f_{\sigma,D_l}(x), \sigma > 0, \}, f_{\sigma,D_l} = \mathrm{sgn}(\frac{1}{l}\sum_{i=1}^{l} y_i g(x - x_i, \sigma))$$

  - The generalization error

$$R(f_{\sigma,D_l}) = \int_Z | y - f_{\sigma,D_l}(x)) | \, d\tilde{F}(x, y).$$

  - The learning implementation scheme (the procedure of finding σ_E^*) (Is it a learning algorithm?)

$$f_{\sigma_E^*,D_l} = \mathrm{sgn}(\frac{1}{l}\sum_{i=1}^{l} y_i g(x - x_i, \sigma_E^*))$$

4. Theory: Visual classification machine

- Learning theory of VCM
  - How can the generalization performance of VCM be controlled (what is the learning principle)?
  - If is it convergent? (If it is a learning algorithm?)

Key is to develop a rigorous upper bound estimation on

$$P\left\{\left|R(f_{\sigma,D_l})-OPT_F(H)\right|\right\}$$

and estimate

$$P\left\{\left|R(L(D_l))-OPT_F(H)\right|<\varepsilon\right\}.$$

4. Theory: Visual classification machine

Theorem 1. *For any* $\sigma > 0$, *we have* $\forall\ \delta \in (0,1), \forall\ \varepsilon > 0$,

$$P^l\{|R(f_{\sigma,\mathcal{D}_l}) - OPT_F(H)|\} \leq \frac{1}{2}(\delta + P^X\{0 < |E_y(x)P(x)| < Bound(\varepsilon,\delta,\sigma,l))\},$$

where

$$Bound(\varepsilon,\delta,\sigma,l) = c_1\frac{a(\delta)+b(\varepsilon)}{l^{\frac{1}{2}}(\sigma)^n} + c_1\varepsilon + c_2\frac{\sqrt{2\pi}a(\delta)}{l^{\frac{1}{2}}} + \sqrt{2\pi}c_2\left(\varepsilon + b(\varepsilon)\right)\sigma^{n+1};$$

$$c_1^{-1} = \inf_{\sigma\in[0,1]}\int_X g(z-x,\sigma)dx > 0, c_2^{-1} = m(X)g(z-x,1);$$

$$a(\delta) = \sqrt{\frac{2\ln\frac{2}{\delta}}{(\sqrt{2\pi})^{2n}}} > 0, b(\varepsilon) = \frac{1}{(2\pi)^{\frac{n}{2}}c_\varepsilon^{(n+1)}}(n + \frac{2B^2}{c_\varepsilon^{2(n+1)}}) > 0.$$

A This theorem shows that to maximize the generalization of the machine is equivalently to minimize $Bound(\varepsilon,\delta,\sigma,l)$

4. Theory: Visual classification machine

Theorem 2. *The minimum of* $Bound(\varepsilon, \delta, \sigma, l)$ *is*

$$\min_{\sigma>0} Bound(\varepsilon, \delta, \sigma, l) = a_1 l^{-\frac{n+1}{4n+2}} + a_2$$

which is attained at the scale

$$\sigma^*_{l,\varepsilon,\delta} = \left(\frac{c_1 n \left(a(\delta) + b(\varepsilon)\right)}{\sqrt{2\pi} c_2 (n+1) \left(\varepsilon + b(\varepsilon)\right)} \right)^{\frac{1}{2n+1}} \frac{1}{l^{\frac{1}{4n+2}}}$$

where

$$a_1 = \sqrt[2n+1]{\left(c_1 \frac{a(\delta) + b(\varepsilon)}{(n+1)} \right)^{n+1} \left(\frac{\sqrt{2\pi} c_2 \left(\varepsilon + b(\varepsilon)\right)}{n} \right)^{n}},$$

$$a_2 = c_1 \varepsilon + c_2 \frac{\sqrt{2\pi} a(\delta)}{l^{\frac{1}{2}}}$$

and n *is the dimension of input space of learning machine.*

A VCA is just designed to approximate $\sigma^*_{l,\varepsilon,\delta}$ here. This reveals the learning principle behind VCA and explain why VCA has strong generalization capability.

4. Theory: Visual classification machine

Theorem 3. *Let $f_{\sigma,\mathcal{D}_l}(x)$ denote the VCA learning function and σ_E^* denote the optimal scale found by 5-cross validation, then we have $\forall\ \varepsilon > 0$ and $\delta \in (0,1)$, there is an integer $l(\varepsilon.\delta)$ such that if $l > l(\varepsilon.\delta)$*

$$P^l\left(\left(R\left(f_{\sigma_E^*,\mathcal{D}_l}(x)\right) - OPT_F(H)\right) \geq \varepsilon\right) < \delta$$

where P^l is over the i.i.d. elements of $\mathcal{D}_l$ generated from the probability distribution $F^l(x,y)$.

A This theorem shows that the VCA is a learning algorithm. Consequently, a learning theory of VCM is established.

5. Concluding remarks

- The existing approaches for classification has mainly been aimed to exploring the intrinsic structure of dataset, less or no emphasis paid on simulating human sensation and perception. We have initiated an approach for classification based on human visual sensation and perception principle (The core idea is to model the blurring effect of lateral retinal interconnections based on scale space theory). The preliminary simulations have demonstrated that the new approach potentially is encouraging and very useful.
- The main advantages of the new approach are its very high efficiency and excellent scalibility. It very often brings a significant reduction of computation effort without loss of prediction capability, especially compared with the prevalently adopted SVM approach.
- The theoretical foundations of VCA, Visual learning theory, have been developed, which reveals that (1) VCA attains its high generalization performance via minimizing the upper error bound between actual and optimal risks (learning principle); (2) VCA is a learning algorithm.

Thanks *!*