Ferdowsi University of Mashhad

Loss Function in Clustering

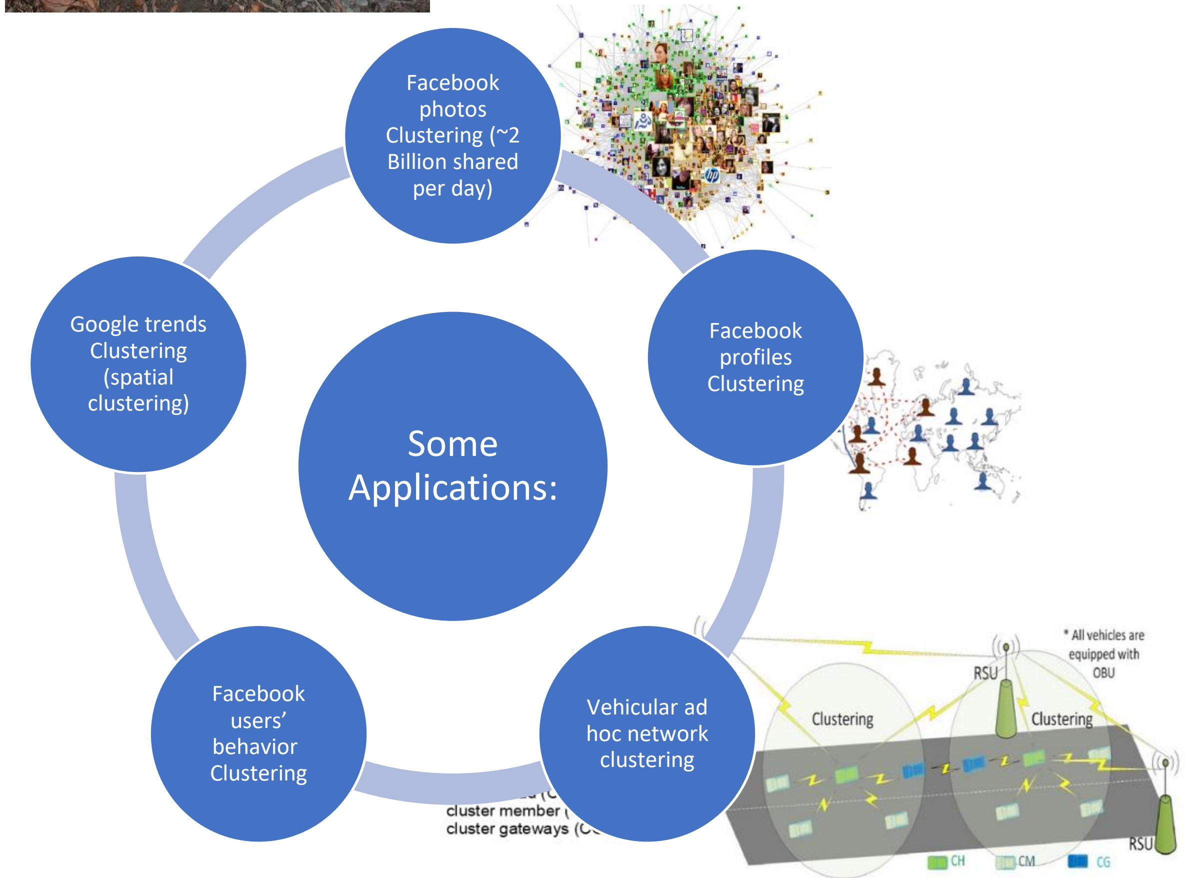
Hadi Sadoghi Yazdi
Department of Computer Engineering
Ferdowsi University of Mashhad

Table of Contents

Data Visualization-based Clustering

A few examples of clustering

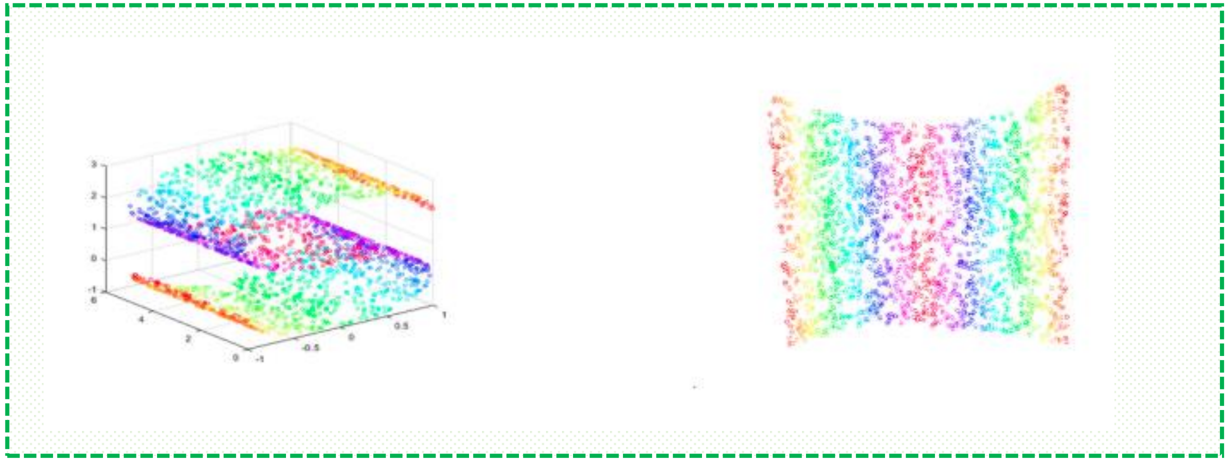
Clustering: a group of similar things that are close together.



Data Visualization

- Examples with 3 Features (Attributes)

- S-Curve Dataset



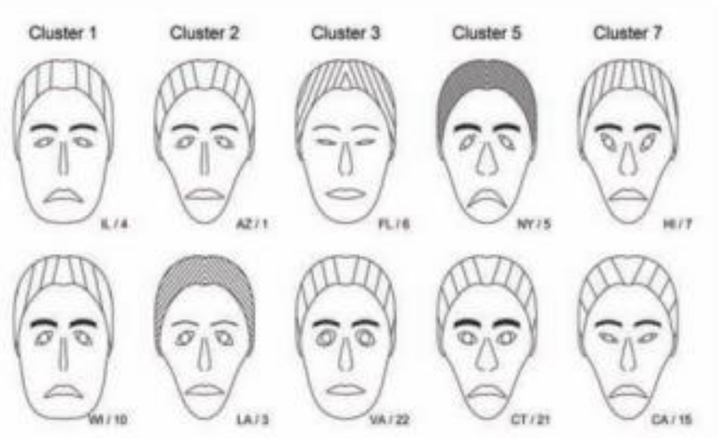
Data Visualization and clustering

Chernoff face

Face features

option	face feature
isize(*exp*)	eye size
iangle(*exp*)	eye angle
ihor(*exp*)	eye horizontal position
ivert(*exp*)	eye vertical position
psize(*exp*)	pupil size
ppos(*exp*)	pupil position
bcurv(*exp*)	brow curvature
bdens(*exp*)	brow density
bhor(*exp*)	brow horizontal position
bvert(*exp*)	brow vertical position
fline(*exp*)	face line
hupper(*exp*)	hair upper line
hlower(*exp*)	hair lower line
hdark(*exp*)	hair darkness
hslant(*exp*)	hair shading slant
nose(*exp*)	nose line
msize(*exp*)	mouth size
mcurv(*exp*)	mouth curvature

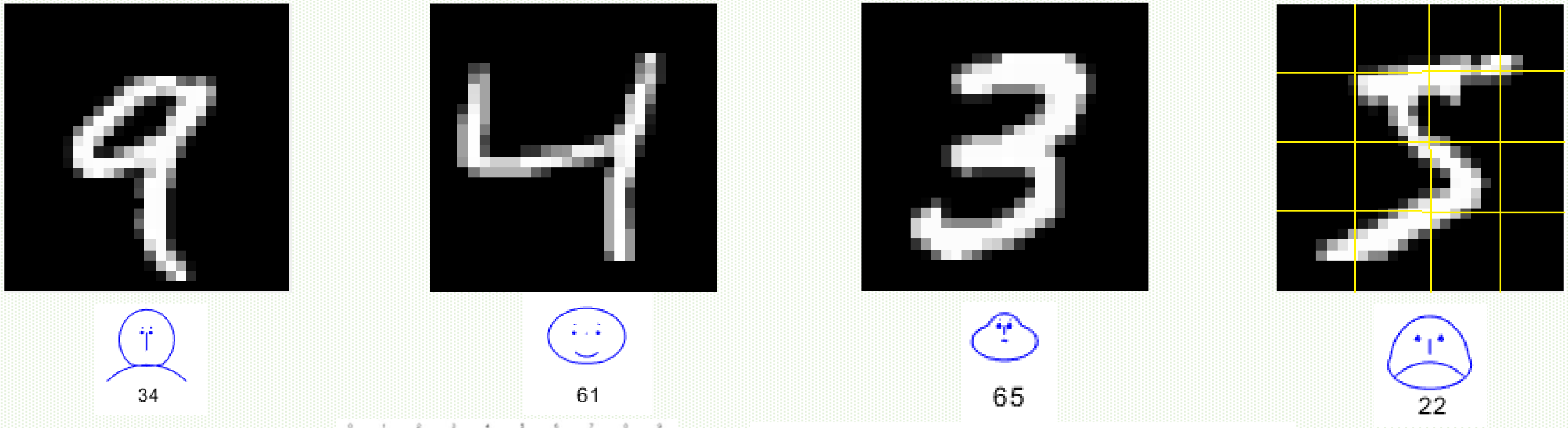
Clustering with Human

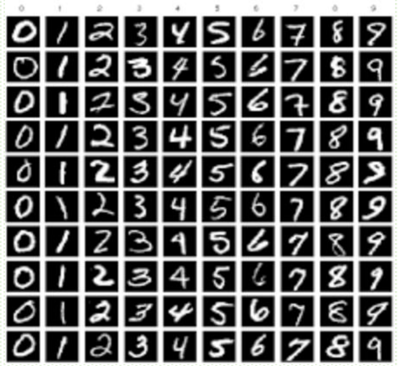


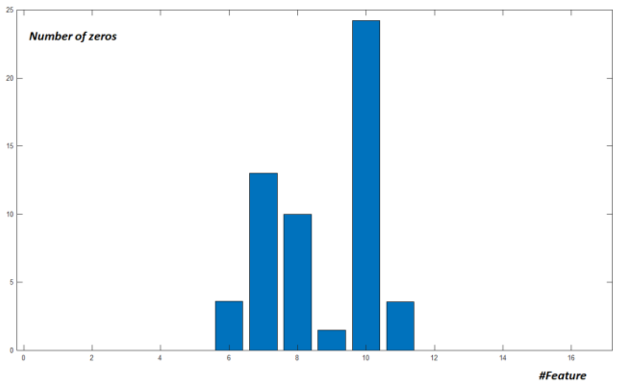
Data Visualization and clustering

Simple Example: Mnist

- Feature Extraction (Zonning), Visualization of 16 extracted features



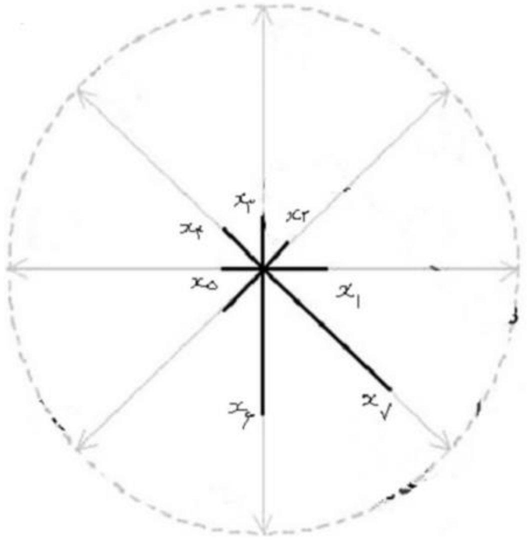




Data Visualization and clustering

- **Star Coordinate**

$$X = \sum_{i=0}^{d-1} x_i \cos\theta_i\,, \qquad Y = \sum_{i=0}^{d-1} x_i \sin\theta_i\;, \qquad \theta_i = \frac{2\pi i}{d}$$



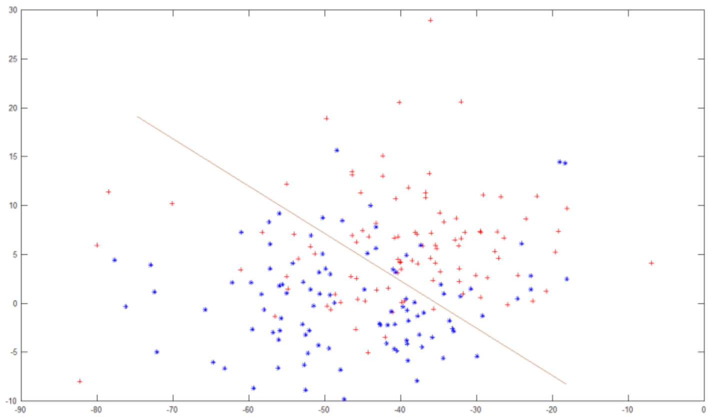
Example: 8 features

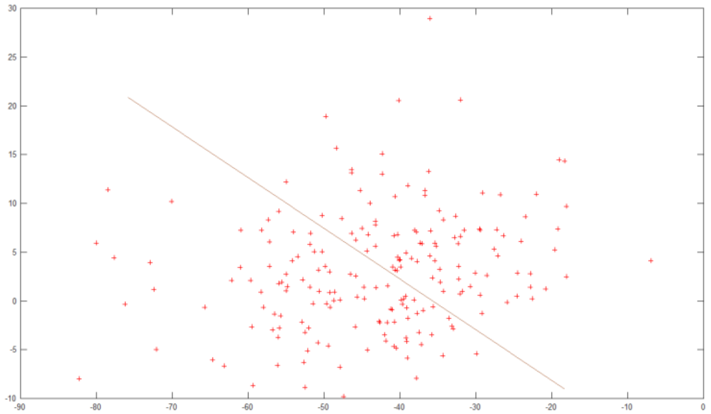
$$P = (x_1, x_2, \cdots, x_8)^T$$

$$X = x_1 \cos 0 + x_2 \cos\frac{\pi}{4} + x_3 \cos\frac{\pi}{2} + \cdots + x_8 \cos\frac{2\pi \times 7}{8}$$

$$Y = x_1 \sin 0 + x_2 \sin\frac{\pi}{4} + x_3 \sin\frac{\pi}{2} + \cdots + x_8 \sin\frac{2\pi \times 7}{8}$$

For: 9, and 4



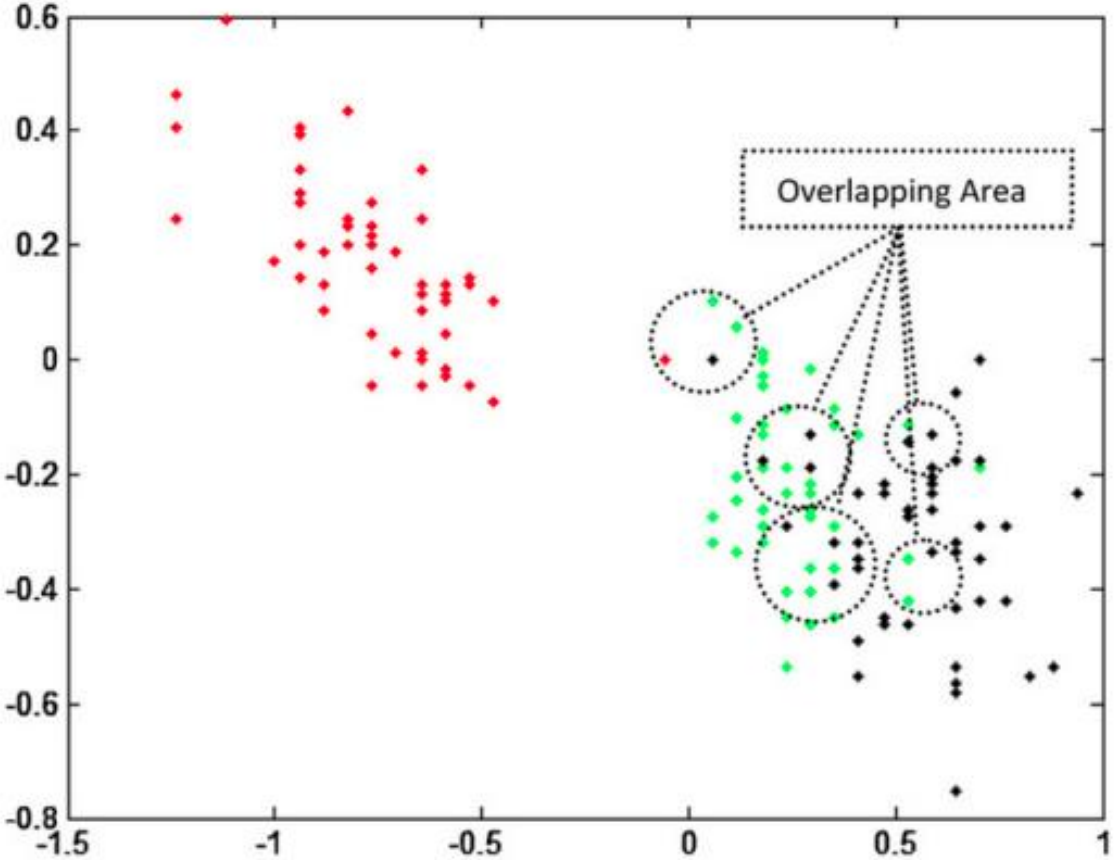


9,4

Data Visualization and clustering

Star Coordinate: Iris (4 dim)

- Using Semi-supervised Learning, Rotate axes with some labeled data. Fisher Criteria, Classification Rate and so on are used for finding best angle.



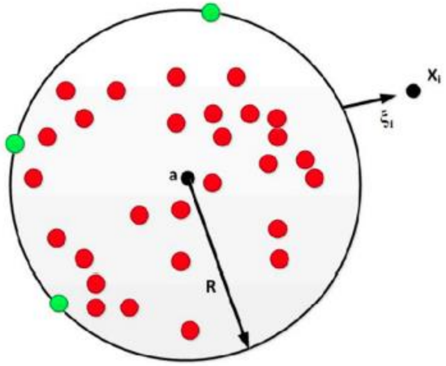
et.al, sadoghi, ‘Discernible visualization of high dimensional data using label information’ , Applied Soft Computing, Volume 27, 2015

Index and Loss Function

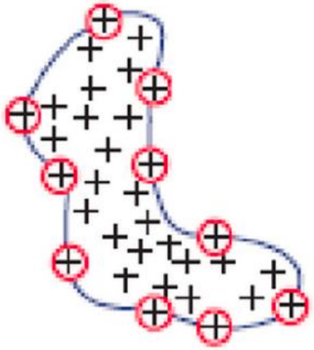
One Cluster of Data

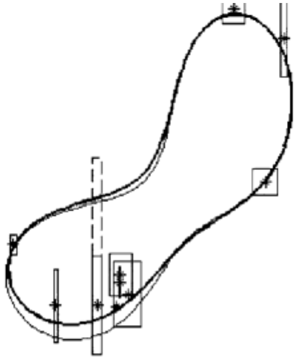
Data Description

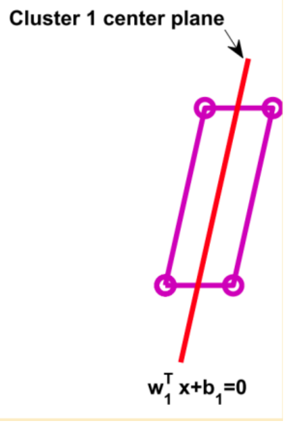
center



Boundary







Some Examples:

- Modelling with Lack of described data [1]
- Modelling of Huge Data via EM [3]
- Behavioral Description in classifier fusion[2]
- Inaccurate data description [4]
- For finance[5]
- …

[1] et al**, Hadi Sadoghi Yazdi**, ‘Prediction of liquefaction potential based on CPT up-sampling’, Computers & Geosciences, 2012

[2] et al**, Hadi Sadoghi Yazdi**, ‘Creating and measuring diversity in multiple classifier systems using support vector data description’, Applied Soft Computing, 2011

[3]] et al**, Hadi Sadoghi Yazdi**, ‘Sparsity-aware support vector data description reinforced by expectation maximization’, Expert System, 2021

[4] et al, **Hadi Sadoghi Yazdi**, ‘ An extension to fuzzy support vector data description’, Pattern Analysis and application 2012

[5] et al**, Hadi Sadoghi Yazdi**, ‘ An Empirical Modeling of Companies Using Support Vector Data Description’, 2010

Cognition of Loss Function

- The first question is why do we think of using the loss function?

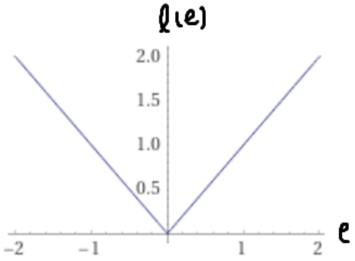
  - ✓ Facts like death, poverty, pain, Short life, Fear, and so on, are painful losses.
  - ✓ Losses in buying and selling, stock market, marriage, partnership, living.

- Is loss function a measure of the error magnitude?

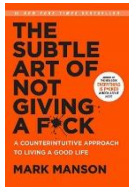
magnification error

Error correction

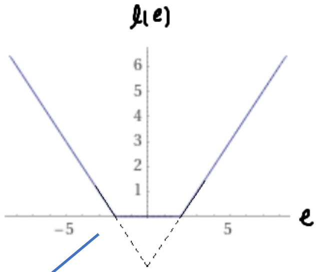
Different kinds of Loss Function



+



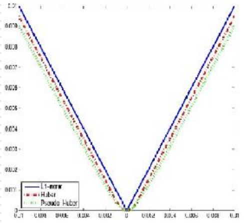
$max(|e| - \epsilon, 0)$



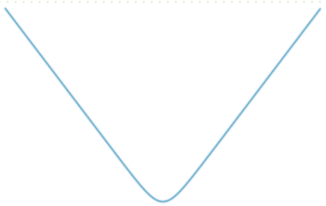
Differentiable

Pseudo-Huber

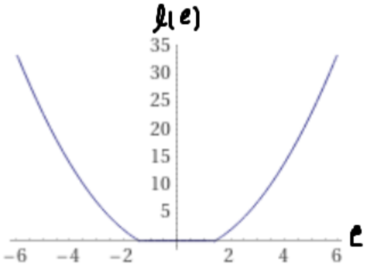
$$L_\delta(a) = \delta^2 \left(\sqrt{1 + (a/\delta)^2} - 1\right)$$



LnCosh



$max(e^2 - \epsilon, 0)$



For Example, Where is the LnCosh formed?

- Abs(error)--->Differentiable→ LnCosh

- Sign Least Mean Square [1]

(simple type of Stochastic gradient descend)

$$w(n+1) = w(n) - \mu \frac{\mathrm{d}E(w)}{\mathrm{d}w},$$
$$w(n+1) = w(n) + \mu sign(e(n)) x^T(n),$$
$$sign(e(n)) \approx tanh(e(n))$$

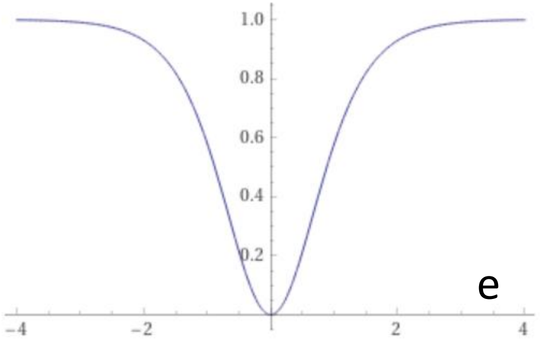
$$\frac{\mathrm{d}E(w)}{\mathrm{d}w} = -tanh(e(n)) = -tanh(d(n) - x^T(n)w(n)).$$

$$E(w) = \int -tanh(d(n) - x^T(n)w(n))\mathrm{d}w$$
$$= \frac{x^T(n)}{x^T(n)} ln(cosh(d(n) - x^T(n)w(n))) = ln(cosh(e(n))).$$

- [2]

Robustness of LnCosh

$$1 - e^{-\alpha lncosh(e)}$$



[1] Verhoeckx, N., van den Elzen, H., Snijders, F., & van Gerwen, P. (**1979**). Digital echo cancellation for baseband data transmission. IEEE Transactions on Acoustics, Speech, and Signal Processing, 27(6), 768–781.

[2] et al**, Hadi Sadoghi Yazdi**, ‘ Robust classification via clipping-based kernel recursive least lncosh of error,’ Expert Systems With Applications 2022.

Pinball Loss Function

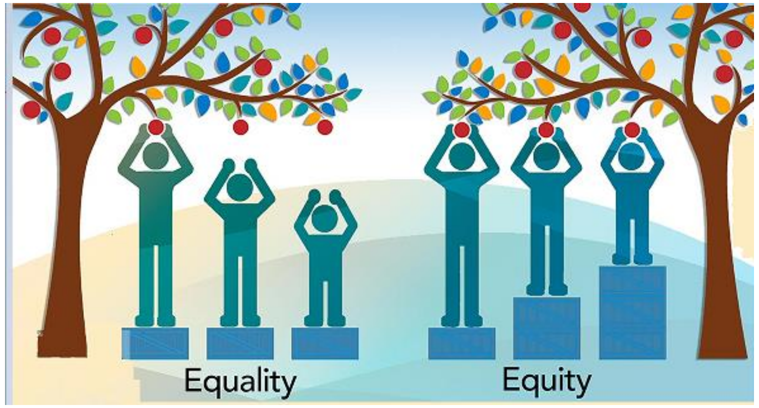
- Motivation:
  - ✓ over-prediction, under-prediction
  - ✓ non-normal errors

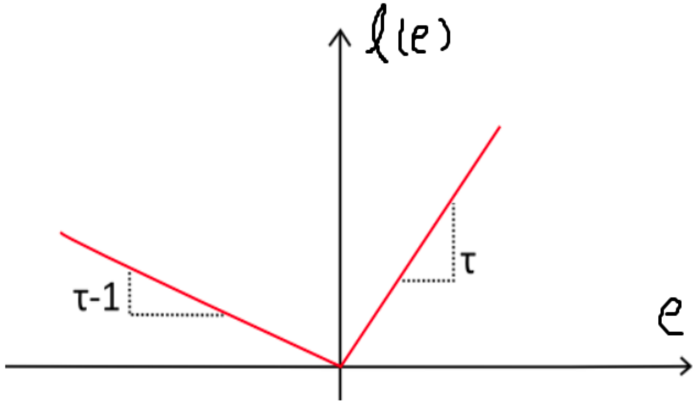
$\varepsilon - insensitive$ Pinball loss function

$$\tilde{e} = e - (\widehat{X}_i \omega_i + b_i e)$$

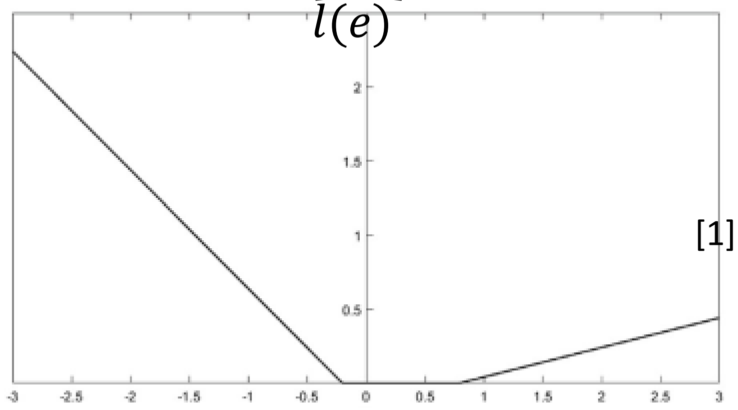
$$\min_{X_i, \omega_i, b_i} \|X_i \omega_i + b_i e\|^2 + c \sum_{i=1}^{n} l(\widehat{\tilde{e}}) \quad \text{s.t. } ||w_i||^2 = 1$$

$$y = \arg\min_i \left\{ |w_i^\top x + b_i|, \quad i = 1, \ldots, k \right\}$$





$\widehat{l(e)}$



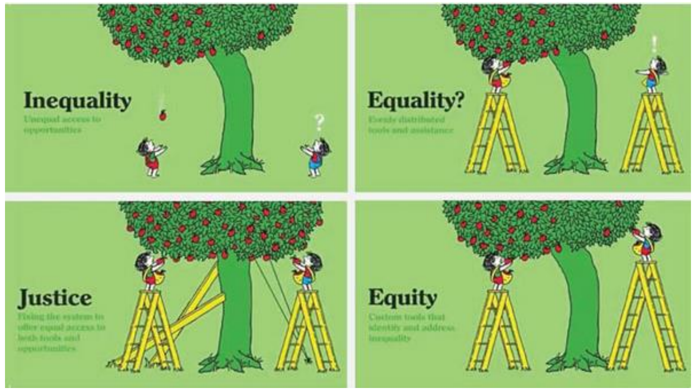
[1]

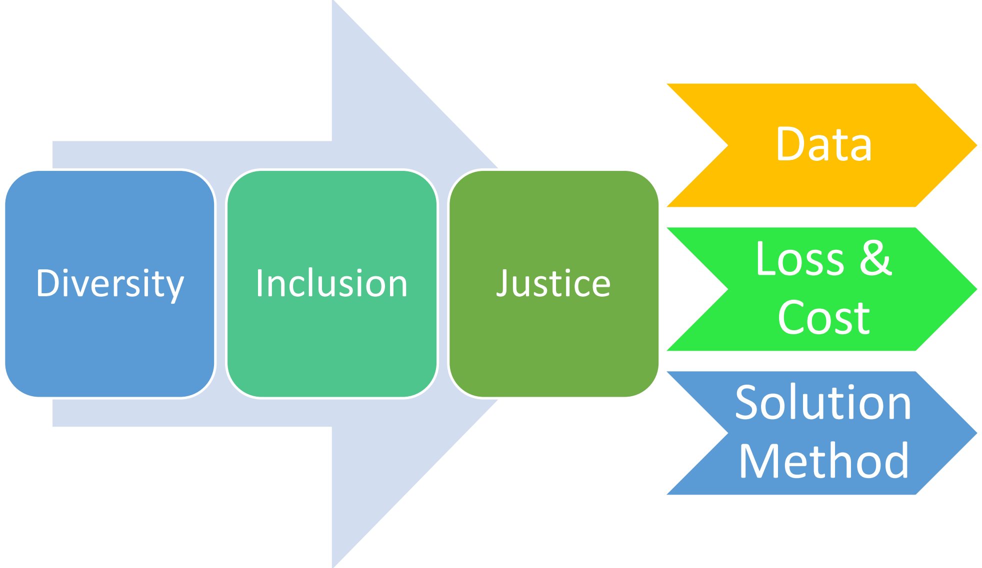
[1] Pritam Anand, 'A new asymmetric ϵ-insensitive pinball loss function based support vector quantile regression model,' Applied Soft Computing, 2020

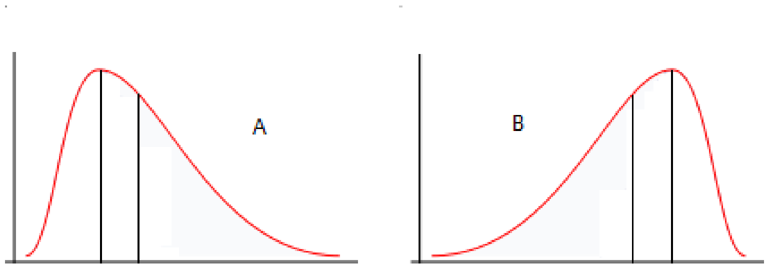
[2] M. Tanveer, et. Al, 'Sparse Twin Support Vector Clustering Using Pinball Loss,' IEEE Journal of Biomedical and Health Informatics 2021

Diversity, Equity, Justice

- Weighting based on decision profile [1]
- Diversity needs robust loss function [2]
- Equity or Fairness leads to bias [3]
- Diversity is good [4]
- features that closely approximate the non-sensitive features[5]





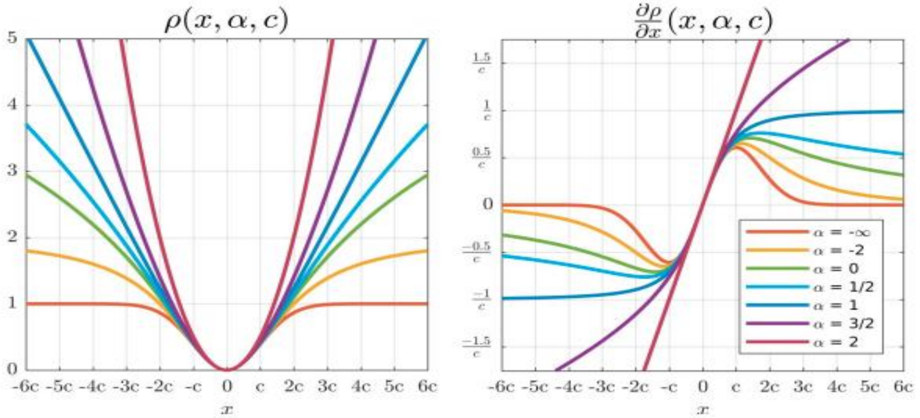


[1] et al, Sadoghi ,'Making Diversity Enhancement Based on Multiple Classifier System by Weight Tuning,' Neural Process Lett (2012)
[2] et al, Sadoghi, 'Diversity-based diffusion robust RLS using adaptive forgetting factor,' Signal Processing 2020.
[3] Onur Köksoy, et.al,' A new right-skewed loss function in process risk assessment,' European Journal of Industrial Engineering, 2019
[4] et al, Sadoghi ,'Creating and measuring diversity in multiple classifier systems using support vector data description,' Applied soft computing, 2011.
[5] Steffen Grünewälder, et.al, 'Oblivious Data for Fairness with Kernels,' jmlr, 2021.

Loss function, Ensemble Learning

$$\ell(x,\alpha,c)=\frac{|\alpha-2|}{\alpha}\left(\left(\frac{\left(\frac{x}{c}\right)^2}{|\alpha-2|}+1\right)^{\frac{\alpha}{2}}-1\right),\quad \xrightarrow{[1]}\quad \rho(x,\alpha,c)=\begin{cases}\frac{1}{2}\left(x/c\right)^2 & \text{if } \alpha=2\\ \log\left(\frac{1}{2}\left(x/c\right)^2+1\right) & \text{if } \alpha=0\\ 1-\exp\left(-\frac{1}{2}\left(x/c\right)^2\right) & \text{if } \alpha=-\infty\\ \frac{|\alpha-2|}{\alpha}\left(\left(\frac{\left(x/c\right)^2}{|\alpha-2|}+1\right)^{\alpha/2}-1\right) & \text{otherwise}\end{cases}$$

$$R\{L(\gamma,\theta,\theta^*)\}=E\left\{\sum_{k=1}^{m}\gamma_k l_k(\theta,\theta^*)\right\}\quad [2]$$



[1] Jonathan T. Barronm, 'A General and Adaptive Robust Loss Function,' 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)
[2] et al, Sadoghi ,'RELF: Robust Regression Extended with Ensemble Loss Function,' Applied Intelligence, 2019

Index and Loss Function

$$J(\mu) = \sum_i e_i^2$$

$$e_i = (x_i - \mu)$$

Square Loss

$$J(\mu) = (x_1 - \mu)^2 + (x_2 - \mu)^2 + \cdots + (x_n - \mu)^2$$

$$\frac{\partial J}{\partial \mu} = 0 \Rightarrow \left(\sum_{i=1}^{n} 2(-1)\,(x_i - \mu) \right) = 0 \Rightarrow \mu = \frac{1}{n} \sum_i x_i$$

$$\max_{\mu} J = \sum_{i=1}^{n} 1 - exp(-\eta(x_i - \mu)^2)$$

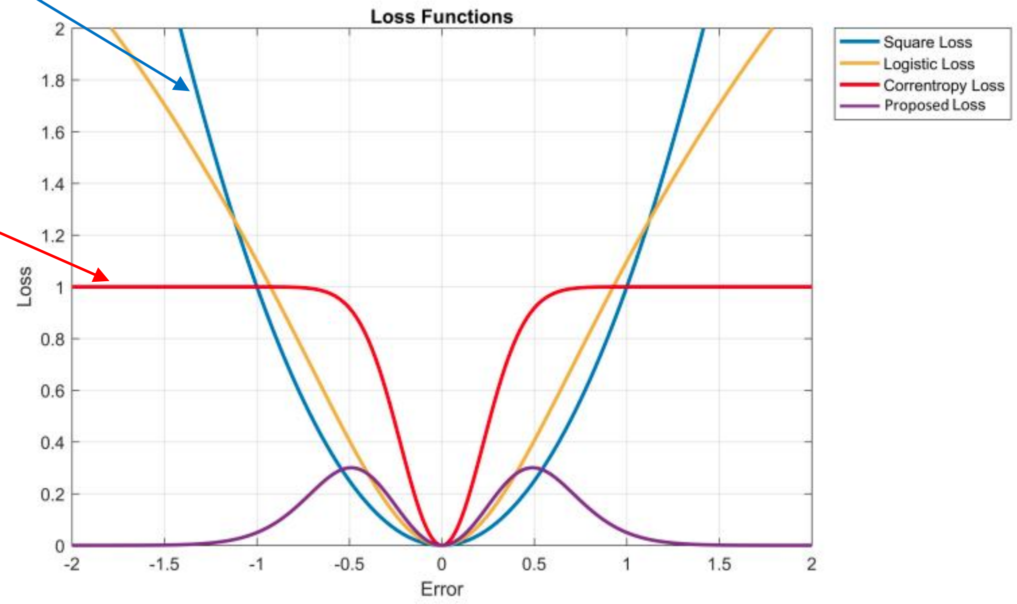
Correntropy Loss

$$\Rightarrow \mu = \frac{\sum_{i=1}^{n} x_i exp(-\eta(x_i - \mu)^2)}{\sum_{i=1}^{n} exp(-\eta(x_i - \mu)^2)}$$

Nature Index

Bird Index

Old Vehicle Index



$Index\ by\ \epsilon - insensitive\ absolute\ loss$

$$\min_{\mu} \sum_{i=1}^{n} max(|x_i - \mu| - \epsilon, 0)$$

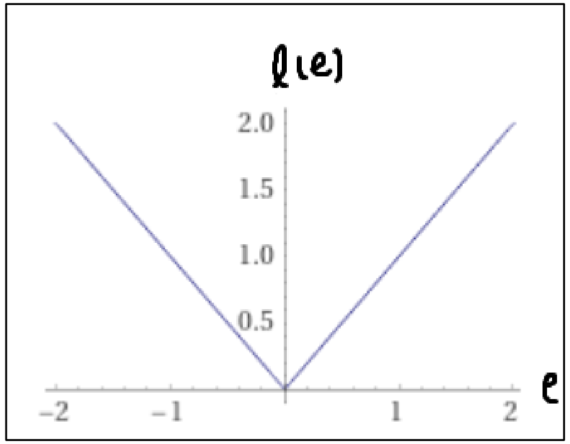
ξ_i

$$\min_{\mu} \sum_{i=1}^{n} \xi_i$$

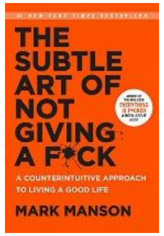
s.t.

$$\xi_i \geq 0 \qquad i = 1, \dots n$$

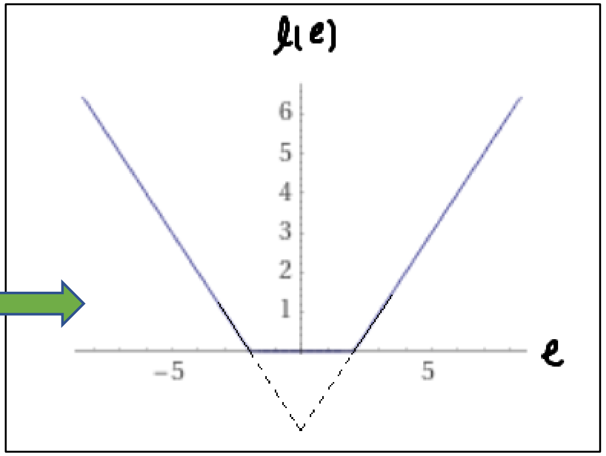
$$\xi_i \geq |x_i - \mu| \qquad i = 1, \dots n$$



+

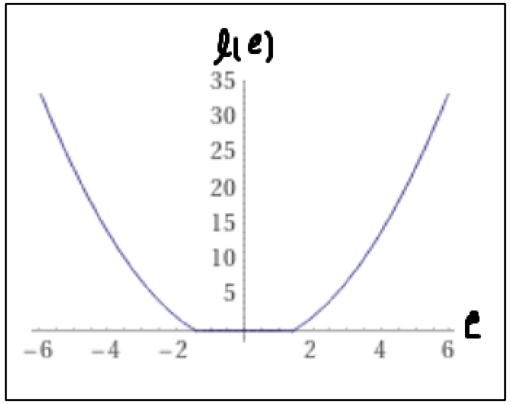


$max(|e| - \epsilon, 0)$



$Index\ by\ \epsilon - insensitive\ square\ loss$

$$\min_{\mu} \sum_{i=1}^{n} max(\|x_i - \mu\|^2| - \epsilon, 0)$$



Boundary

$e_i = ||x_i - a||^2 - R^2$ → ε -insensitive

$$l(e_i) = \max(0, ||x_i - a||^2 - R^2)$$

$$\arg min \sum l(e_i) + Regularization\ term$$

$for\ example\ Regularization\ term\ is\ minimum\ Radius$
$\min R^2$

$$\min \quad R^2 + C \sum_i \xi_i \ ,$$

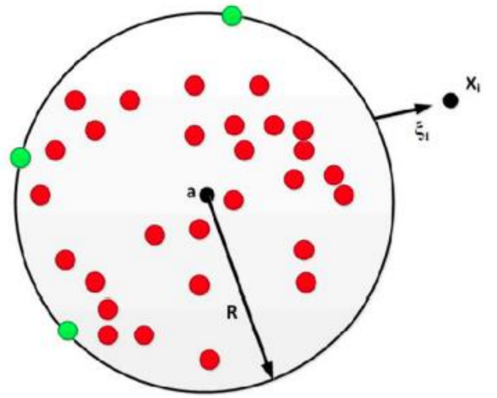
$$\text{s.t.} \ \|x_i - a\|^2 \leq R^2 + \xi_i, \forall i$$

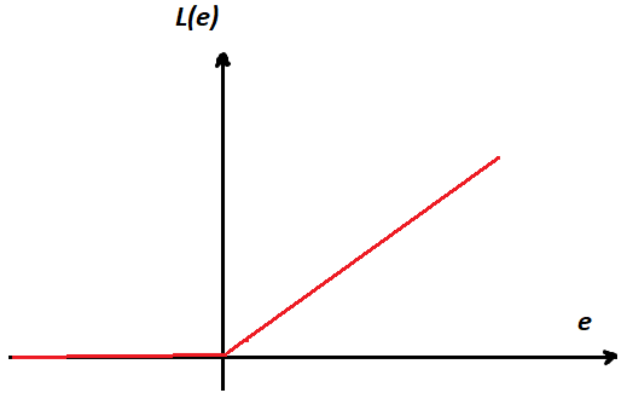
Where

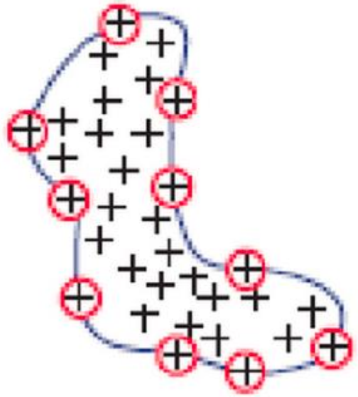
$$x \rightarrow \varphi(x)$$

$$k(x, y) = < \varphi(x), \varphi(y) >$$

Support Vector Data Description





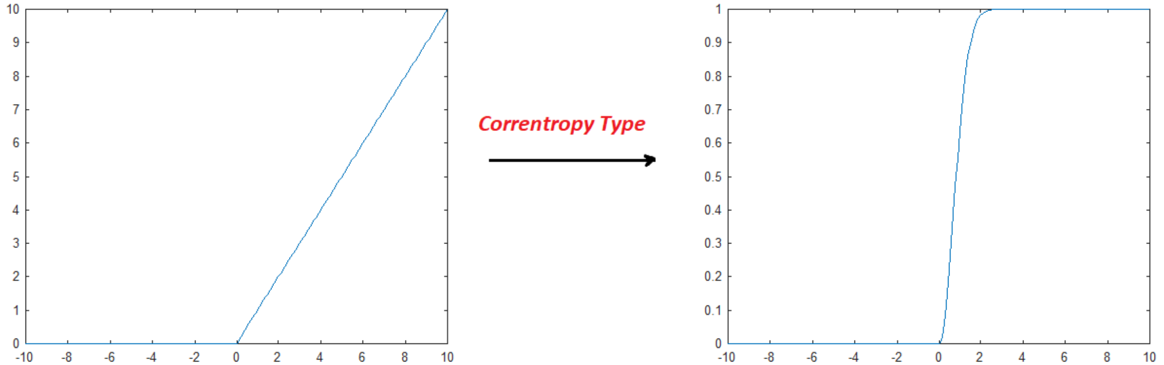


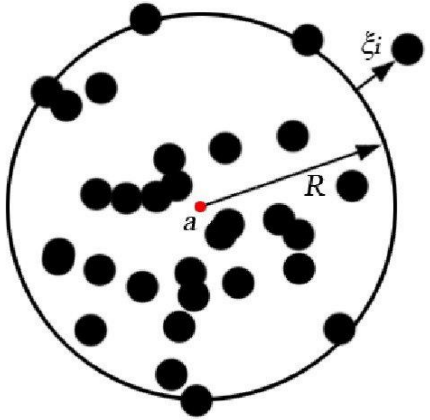
Support Vector Data Description

In [1] correntropy loss: $\sum 1 - e^{-\beta \xi_i} + Regularization\ Term$

In [2] weigthed loss: $\sum w_i \xi_i + Regularization\ Term$

Also another loss function can be applied





Single Sphere

[1] H-J Xing, L-F Li, 'Robust least squares one-class support vector machine', Pattern Recognition Letters 2020
[2] M. Bicegoa, M A.T. Figueiredo, 'Soft clustering using weighted one-class support vector machines', pattern recognition 2009

Center-Plane

$$\text{Center-plane}_i := w_i^\top x + b_i = 0, \quad i = 1, \ldots, k$$

X = (x1, x2, . . . , xm) T

In reference [1]:

- Assuming these m samples belong to k classes

  rest labels into the matrix $\widehat{X}_i$

  ith cluster by $X_i \in R^{m_i \times n}$

$$\min_{w_i, b_i, X_i} \frac{1}{2} ||X_i w_i + b_i e||^2$$
$$\text{s.t. } ||w_i||^2 = 1$$

Square Loss

- Starts from a random initial assignment of the samples

  Then, each sample is relabeled by

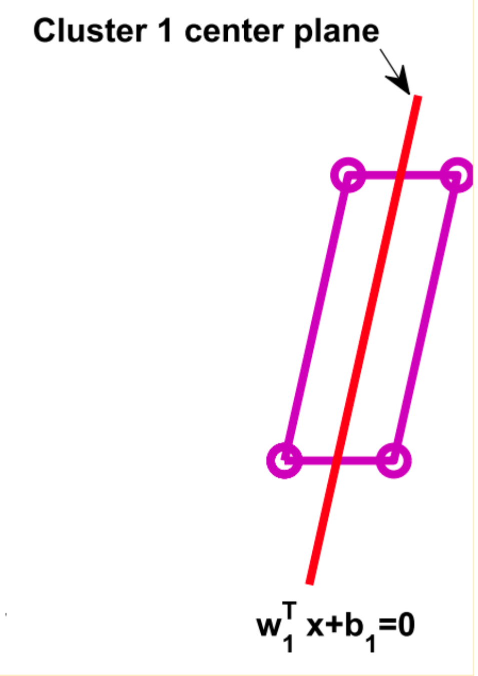
$$y = \arg\min_i \{|w_i^\top x + b_i|, \quad i = 1, \ldots, k\}$$

- Combined with $\epsilon - insensitive\ loss$

Regularization term

$$\min_{X_i, \omega_i, b_i} ||X_i \omega_i + b_i e||^2 + c \sum_{i=1}^{n} max(e - (\widehat{X}_i \omega_i + b_i e) - \epsilon, 0)$$

$$y = \arg\min_i \{|w_i^\top x + b_i|, \quad i = 1, \ldots, k\}$$



$$\min_{X_i, \omega_i, b_i} ||X_i \omega_i + b_i e||^2 + c \sum_{i=1}^{n} l(\widehat{\tilde{e}})$$

$$\text{s.t. } ||w_i||^2 = 1$$

and its geometric meaning is clear. For example, when $i = 1$, its objective function makes the data samples in Class 1 proximal to the first class center plane $w_1^\top x + b_1 = 0$, while the constraints make the data samples in the rest of the classes have a distance at least 1 from this plane from one side.

[1] Zhen Wang, et. al, 'Twin Support Vector Machine for Clustering,' IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, OCTOBER 2015

Loss-based clustering

Clustering

- **Proposed View**

$$E\{l(x,v)\} = \int_x \int_v l(x,v) f(x,v) dx dv$$

or

$$E\{l(x,v)\} = \frac{1}{nc} \sum_x \sum_v l(x,v) f(x,v)$$

$x \in DataSet, v\ is\ parameter$

$$E\{l(x,v)\} = \sum_{i=1}^{n} \sum_{j=1}^{c} l(x,v) f(x_i|v_j) f(v_j)$$

If we let $l(x,v) = \|x-v\|^2$ then:

$$E\{l(x,v)\} = \sum_{i=1}^{n} \sum_{j=1}^{c} \|x_i - v_j\|^2 u_{ij}^m f(v_j)$$

or if $l(x,v) = 1 - \exp(\frac{-\|x-v\|^2}{2\sigma^2})$ then:

$$E\{l(x,v)\} = \sum_{i=1}^{n} \sum_{j=1}^{c} (1 - \exp(\frac{-\|x_i - v_j\|^2}{2\sigma_j^2})) u_{ij}^m f(v_j)$$

Clustering

- Standard FCM

$$J_m(u,v) = \sum_{i=1}^{n}\sum_{j=1}^{c} u_{ij}^m \|x_i - v_j\|^2$$

$$s.t.$$

$$\sum_{j=1}^{c} u_{ij} = 1 \quad \forall i, \quad 0 < \sum_{i=1}^{n} u_{ij} < n$$

- Another view of error

$$\min_{u,v} \quad J(u,v) = \sum_{i=1}^{n} e_i^2$$

$$s.t.$$

$$e_i^2 = \sum_{j=1}^{c} u_{ij}^m d(x_i, v_j),$$

$$\sum_{j=1}^{c} u_{ij} = 1, \quad i = 1, \cdots, n$$

Robust →

$$\max_{u,v} \quad J(u,v) = \sum_{i=1}^{n} \exp(-\eta_i^2 e_i^2)$$

$$s.t.$$

$$e_i^2 = \sum_{j=1}^{c} u_{ij}^m d(x_i, v_j),$$

$$\sum_{j=1}^{c} u_{ij} = 1, \quad i = 1, 2, \cdots, n$$

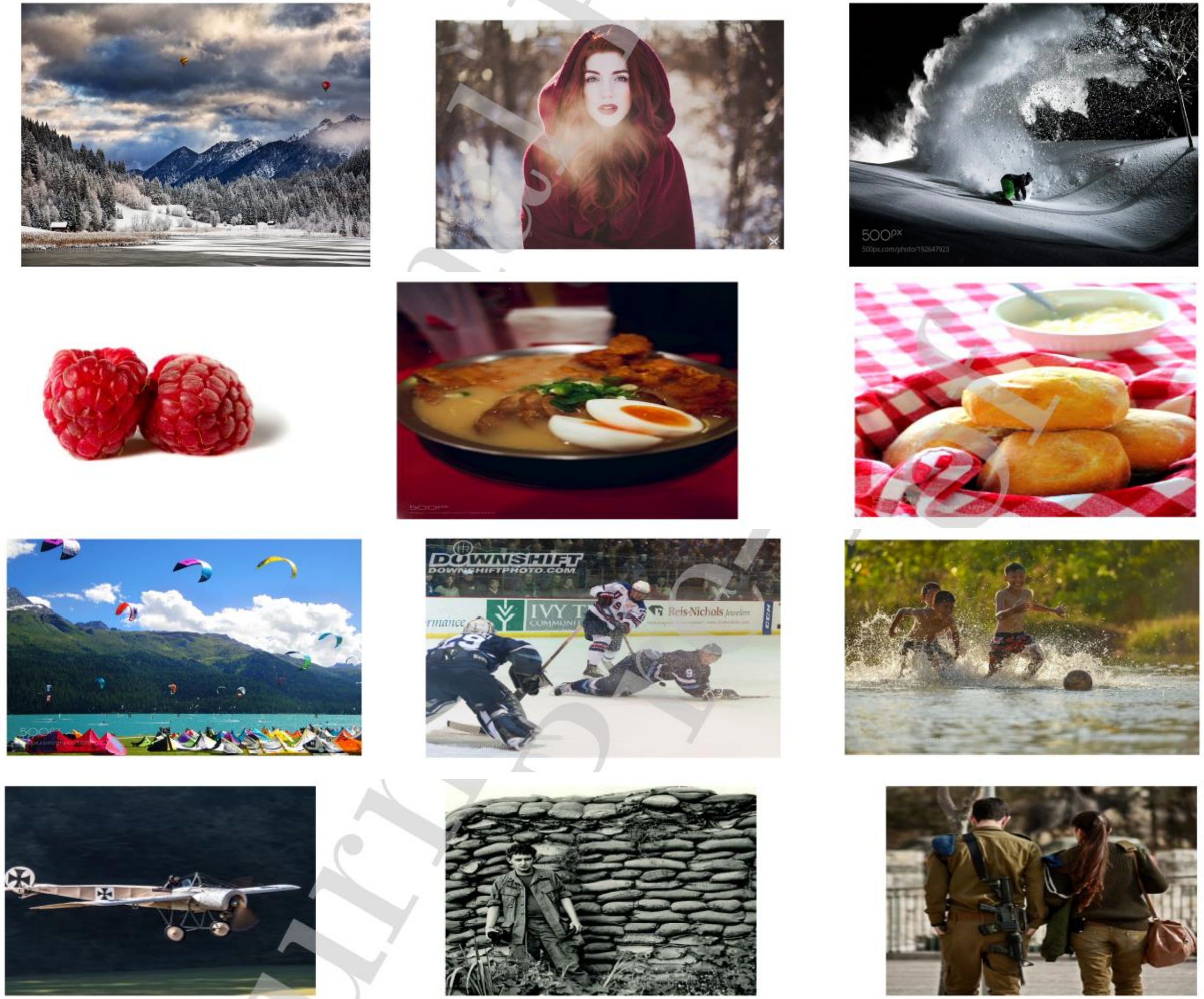
[1] et al, **Hadi Sadoghi Yazdi**, 'Robust Heterogeneous C-means', Applied Soft Computing, 2019

Data Reduction, Concepts index

Example 80,000 images from 500px social media

- Finding **three** images which described 'cold' concept
- Finding **three** images which described 'food' concept
- Finding **three** images which described 'Sport' concept
- Finding **three** images which described 'War' concept

four categories: 'sport', 'food', 'cold' and 'war'



take a look at regularization term

$$E\{l(e)\} + \gamma Regularization\ Term:$$

Projection

$$\min_{W,U,M} \sum_{i=1}^{N}\sum_{k=1}^{c} \|W^T x_i - m_k\|_2^2 u_{ik} + \gamma \|U\|_F^2,$$

$$s.t. \quad W^T W = I, \sum_{k=1}^{c} u_{ik} = 1, u_{ik} \geqslant 0$$

Regularization Term

Sparsification of Membership matrix

$$\gamma \sum_{i=1}^{N}\sum_{k=1}^{c} u_{ik} \log u_{ik}$$

$$\sum_{i=1}^{N} \|x_i - WW^T x_i\|_2^2$$

Reduce reconstruction loss

$$\gamma \sum_{i=1}^{n}\sum_{j=1}^{c} \|f_i - f_j\|^2 w_{ij}^2$$

From Laplacian Matrix

Mean Shift Clustering

$$\hat{R}(\boldsymbol{x}) = E_q[l(\boldsymbol{x},\boldsymbol{y})] = \sum_{i=1}^{n} q(\boldsymbol{x}_i)\, l(\boldsymbol{x},\boldsymbol{x}_i)$$

$$q(\boldsymbol{x}_i) = \frac{1}{n\sqrt{2\pi}h}$$

$$l(\boldsymbol{x},\boldsymbol{x}_i) = 1 - g(\boldsymbol{x},\boldsymbol{x}_i) = 1 - \exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{x}_i\|^2}{2h^2}\right)$$

$$G(\boldsymbol{x}) = \frac{1}{n\sqrt{2\pi}h}\sum_{i=1}^{n} \exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{x}_i\|^2}{2h^2}\right)$$

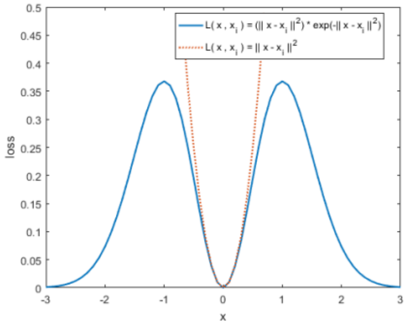
Kernel Density Estimation

$$\max_{\boldsymbol{x}} G(\boldsymbol{x}) = \frac{\partial}{\partial x}\sum_{i=1}^{n} \exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{x}_i\|^2}{2h^2}\right)$$

$$\boldsymbol{x}^{t+1} = \frac{\sum_{i=1}^{n} \boldsymbol{x}_i \exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{x}_i\|^2}{2h^2}\right)}{\sum_{i=1}^{n} \exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{x}_i\|^2}{2h^2}\right)}$$

Mean Shift Clustering

$$\min_{\boldsymbol{x}} \sum_{i=1}^{n} q(\boldsymbol{x}_i)\frac{\|\boldsymbol{x}-\boldsymbol{x}_i\|^2}{2h^2} \exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{x}_i\|^2}{2h^2}\right)$$

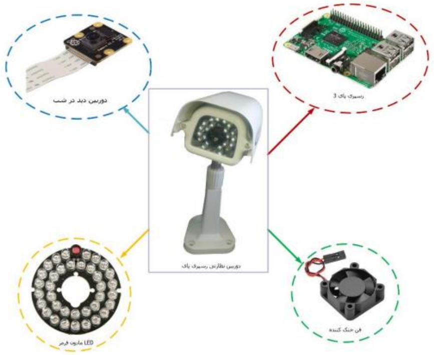


[1] et al, **Hadi Sadoghi Yazdi**, ‘ Crowd analysis using bayesian risk kernel density estimation,’ Engineering Applications of Artificial Intelligence.

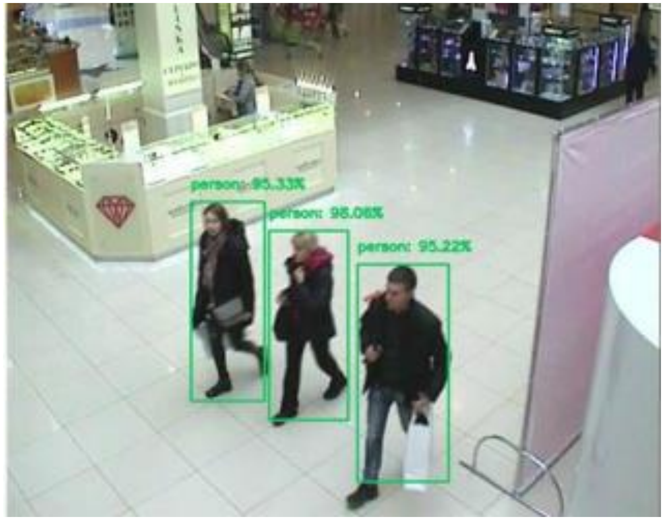
[2] et al, **Hadi Sadoghi Yazdi**, ‘Automated Detection of Region of Interest using Non-Parametric Distribution Based on Bayesian Risk,’ Journal of Machine Vision and Image Processing.

Example for Mean Shift Clustering

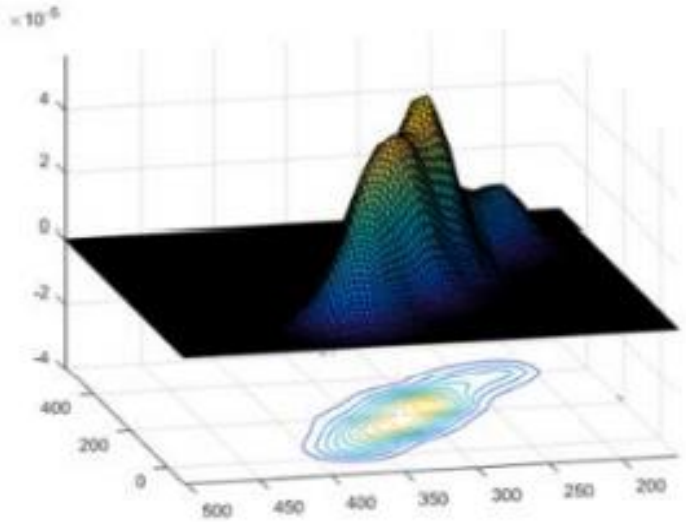
1) Intelligent Camera



2) Detection

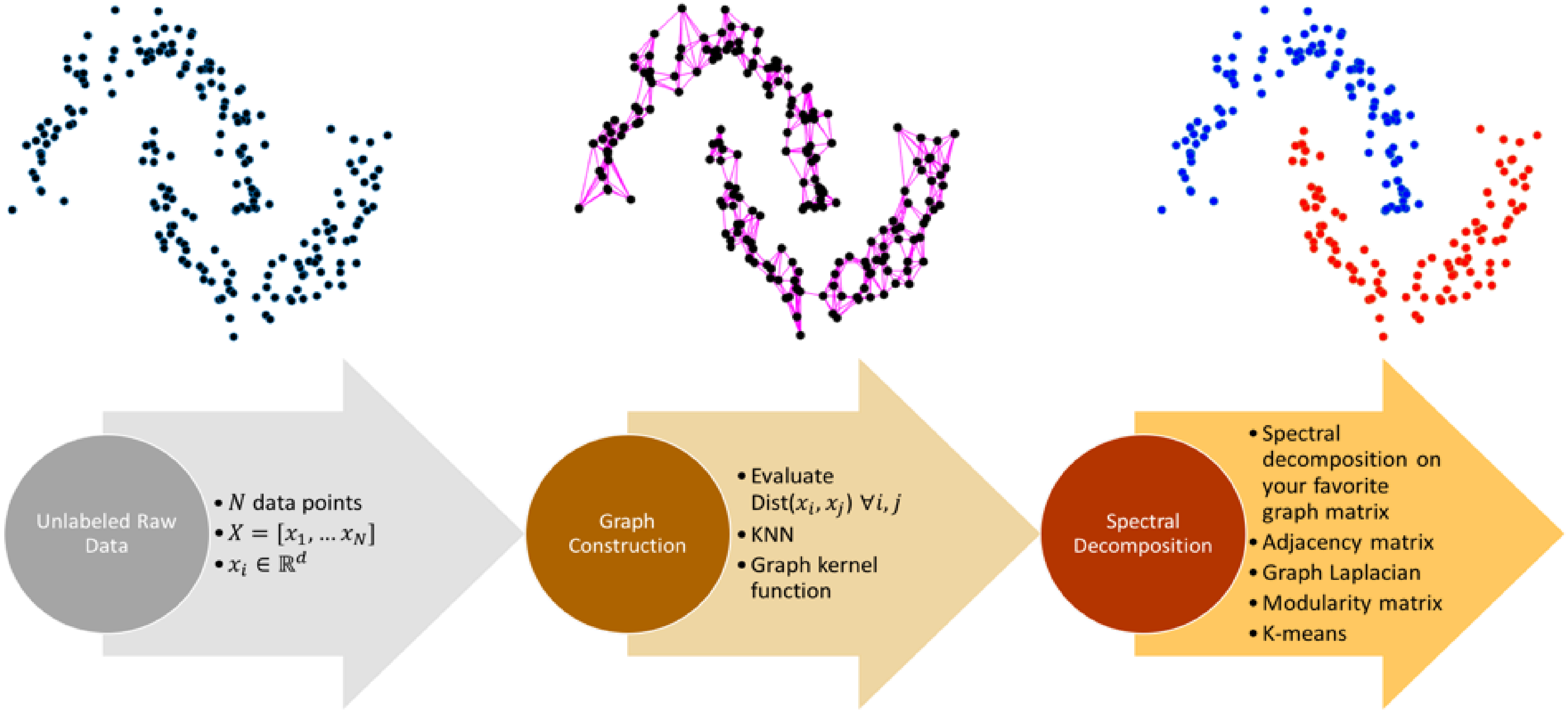


3) Interest Locations by Mean Shift



Spectral Clustering

- Main Steps of Spectral Clustering



Zamiri, M., Bahraini, T. and Yazdi, H.S., 2021. MVDF-RSC: Multi-view data fusion via robust spectral clustering for geo-tagged image tagging. *Expert Systems with Applications*, *173*, p.114657.

Spectral Clustering

- New Loss in Spectral Clustering

$$V_\sigma(X,Y) = E[k(X-Y)], \quad \longrightarrow \quad \widehat{V_\sigma}(X,Y) = \frac{1}{n}\sum_{i=1}^{n} k_\sigma(x_i - y_i),$$

$$\max_{\mathbf{w_i}} \sum_{v=1}^{V}\sum_{i=1}^{N_{tr}} \exp(-\eta_v e_{vi}^2) \quad s.t. \quad \mathbf{1}^T\mathbf{w_i} = 1, \quad \mathbf{w_i} \geqslant 0, \quad \forall i = 1, \dots, N_{tr},$$

$$e_{vi} = \|\mathbf{G_v^i w_i}\|_2, \quad v = \{1, 2, \dots, V\}, \quad i = \{1, \dots, N_{tr}\},$$

$$\mathbf{G_v^i} = Diag(ndist(\mathbf{I_v^1}, \mathbf{I_v^i}), \dots, ndist(\mathbf{I_v^{N_{tr}}}, \mathbf{I_v^i})),$$

✓ Final Cost Function:

where $\mathbf{G_v^i} \in R^{N_{tr}\times N_{tr}}$ is a diagonal matrix having normalized distances between the i^{th} training data and others based on v^{th} view on the diagonal, and is defined as follows:

$$\mathbf{G_v^i} = Diag(ndist(\mathbf{I_v^1}, \mathbf{I_v^i}), \dots, ndist(\mathbf{I_v^{N_{tr}}}, \mathbf{I_v^i})), \tag{5}$$

where $\mathbf{I_v^i}$ denotes the i^{th} column of matrix $\mathbf{I_v}$. Since different views may

$$\max_{\mathbf{w_i},\eta,\mathbf{F}} \sum_{v=1}^{V}\sum_{i=1}^{N_{tr}} \exp(-\eta_v \|\mathbf{G_v^i w_i}\|_2^2) - \lambda \eta^T \mathbf{H} \eta - \gamma Tr(\mathbf{F}^T \mathbf{L_W F})$$

$$s.t. \quad \eta^T \mathbf{1_V} = 1, \mathbf{F}^T\mathbf{F} = \mathbf{I}, \mathbf{1}^T\mathbf{w_i} = 1, \mathbf{w_i} \geqslant 0, \forall i = 1, \dots, N_{tr}.$$

Zamiri, M., Bahraini, T. and Yazdi, H.S., 2021. MVDF-RSC: Multi-view data fusion via robust spectral clustering for geo-tagged image tagging. *Expert Systems with Applications*, *173*, p.114657.

Online Clustering

- Online Index

$$J = E\{l(\mathrm{e})\} \xrightarrow{l(\mathrm{e})=e^2} \hat{J} = \sum_{i=1}^{n} e_i^2 = \sum_{\langle i \rangle} \|x_i - \mu\|^2$$

$$\mu_{new} = \mu_{old} - \lambda \frac{\partial J}{\partial \mu} \qquad SGD$$

$$\mu_k = \mu_{k-1} + 2\lambda \sum_{\langle i \rangle} (x_i - \mu_{k-1})$$

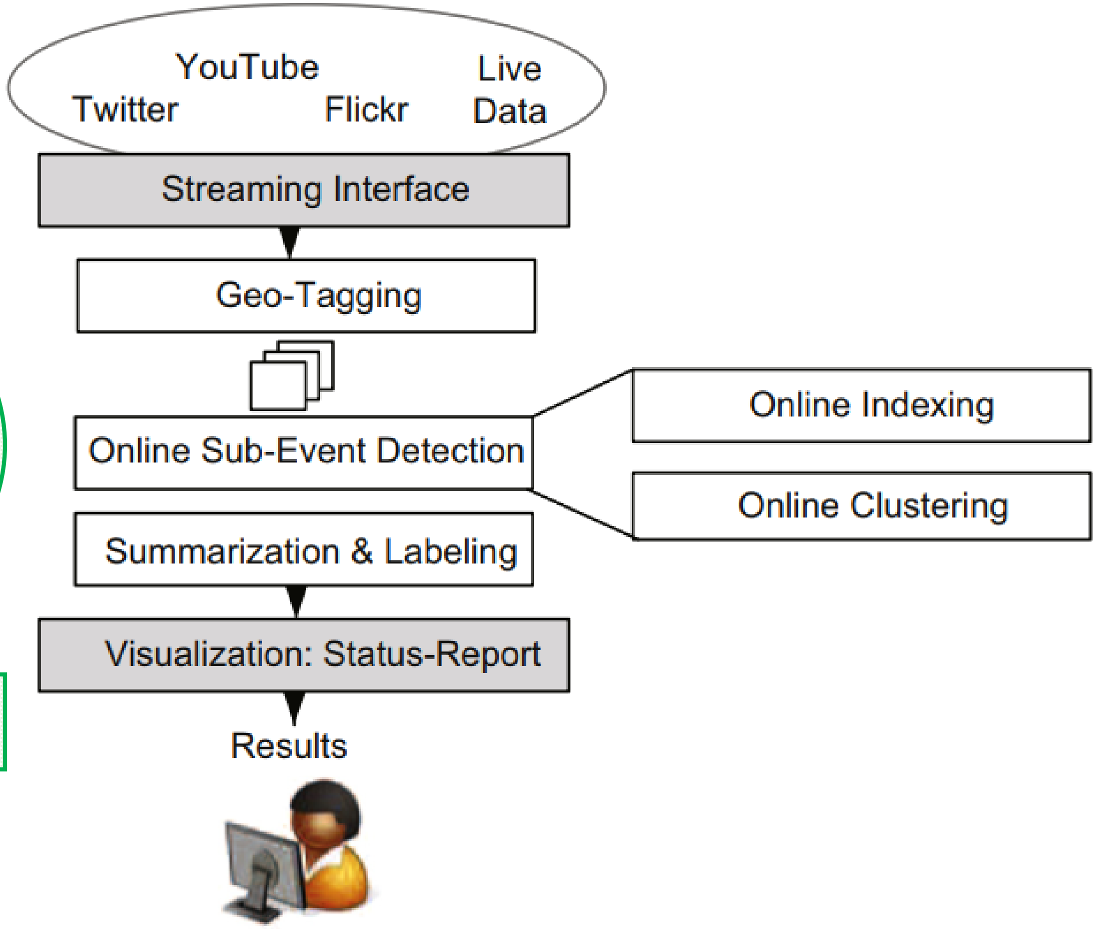


Fig. Multimedia exploration framework.

Pohl, D., Bouchachia, A. and Hellwagner, H., 2016. Online indexing and clustering of social media data for emergency management. *Neurocomputing*, *172*, pp.168-179.

Online Clustering

- Assume new loss function:
  - ✓ Pseudo-Huber loss

$$L_\delta(\mathrm{e}) = \delta^2\left(\sqrt{1+\frac{\mathrm{e}^2}{\delta^2}}-1\right)$$

$$\frac{\partial J}{\partial \mu} = \sum_{i=1}^{n} \frac{1}{2}\left(\frac{-\mu e_i}{\delta^2}\right)\left(1+\frac{e_i^2}{\delta^2}\right)^{-\frac{1}{4}} \Rightarrow \nabla_J = \sum_{i=1}^{n} -(x_i-\mu)\left(1+\frac{\|x_i-\mu\|^2}{\delta^2}\right)^{-\frac{1}{2}}$$

  - ✓ without the coefficients

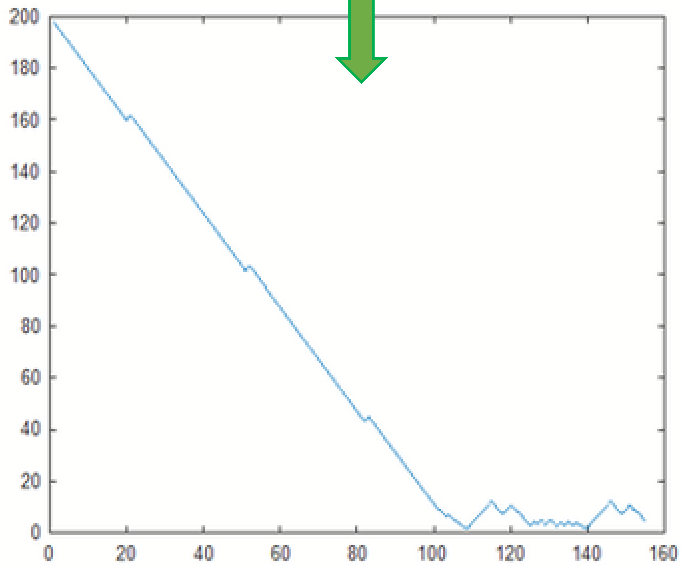
$$\frac{\partial J}{\partial \mu} = \frac{\partial L}{\partial \mu} = -(x_i-\mu)\left(1+\frac{\|x_i-\mu\|^2}{\delta^2}\right)^{-\frac{1}{2}}, \qquad \text{for LMS}$$

Start in 200 as outlier

Convergence to the median value in 4

The output was obtained with 5 iterations on this data

Even with one iteration for large amounts of delta, the answer is obtained

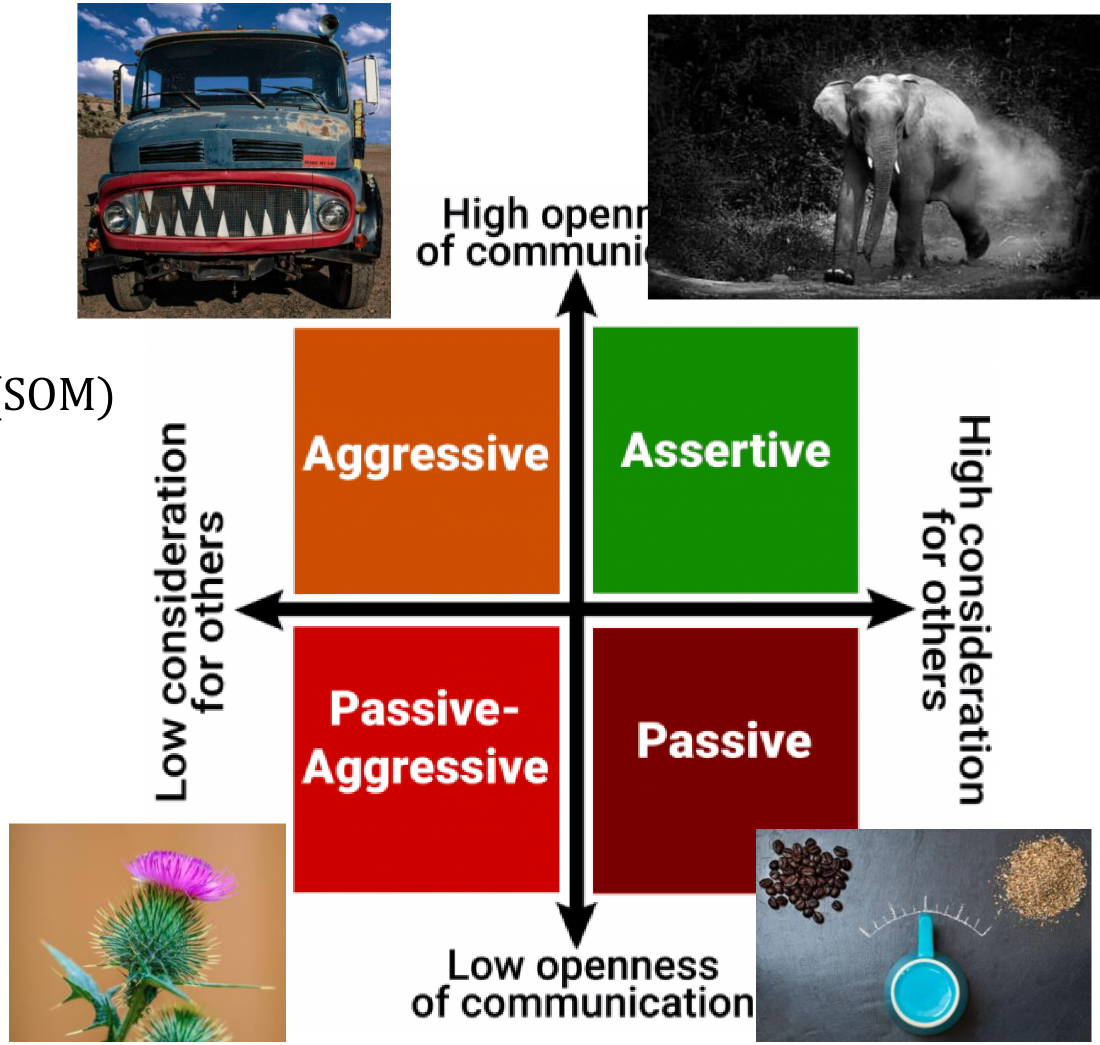


Online Clustering Passive Aggressive

$$\min \frac{1}{2}\|\mu - \mu_{jt}\|^2 + c\xi_i \qquad \text{s.t} \qquad \|x_i - \mu\|^2 \leq \xi_i$$ [1]

$$\mu = \mu_{jt} + (1 - S)(x_i - \mu_{jt})$$

Self Organize Map (SOM)



[1] K. crammer, 'Online Passive-Aggressive Algorithms,' Journal of Machine Learning Research 7 (2006)

Application Examples

Mean Shift Clustering

$$\min_x R_n(x) = \min_x \sum_{i=1}^{n} p(x_i) l(x; x_i) = \min_x \{ \sum_{i=1}^{n} \pi_i \zeta_i \left[1 - K(d(x, x_i; \Sigma_i)) \right] \}.$$

$$\frac{\partial R_n(x)}{\partial x} = 2 \sum_{i=1}^{n} \pi_i \zeta_i K'(d(x, x_i; \Sigma_i)) \times \Sigma_i^{-1} (x - x_i) = 0.$$

$$m_{\Sigma}(x) = \sum_{i=1}^{n} \frac{K'(\frac{\|x - x_i\|^2}{2h^2})}{\sum_{i=1}^{n} K'(\frac{\|x - x_i\|^2}{2h^2})} x_i = \sum_{i=1}^{n} \frac{\exp(-\frac{\|x - x_i\|^2}{2h^2})}{\sum_{i=1}^{n} \exp(-\frac{\|x - x_i\|^2}{2h^2})} x_i .$$

Bahraini, T., Azimpour, P. and Yazdi, H.S., 2021. Modified-mean-shift-based noisy label detection for hyperspectral image classification. Computers & Geosciences, 155, p.104843.

Some Loss functions

Table 1
Different types of loss functions.

	Loss function	Bayesian risk $R_n(x)$	Weighted center of mass $m_{\Sigma}(x)$	Generalized mean shift vector $x^{i+1} = x^i + m_{\Sigma}(x)$
1	$\| x - x_i \|^2$	$\sum_{i=1}^{n} p(x_i) \| x - x_i \|^2$	$\sum_{i=1}^{n} \frac{p(x_i)}{\sum_{i=1}^{n} p(x_i)} x_i$	$\sum_{i=1}^{n} \frac{p(x_i)}{\sum_{i=1}^{n} p(x_i)} x_i - x$
2	$\exp(- \| x - x_i \|^2)$	$\sum_{i=1}^{n} p(x_i) \exp(- \| x - x_i \|^2)$	$\sum_{i=1}^{n} \frac{p(x_i) \exp(- \| x - x_i \|^2)}{\sum_{i=1}^{n} p(x_i) \exp(- \| x - x_i \|^2)} x_i$	$\sum_{i=1}^{n} \frac{p(x_i) \exp(- \| x - x_i \|^2)}{\sum_{i=1}^{n} p(x_i) \exp(- \| x - x_i \|^2)} x_i - x$
3	$1 - \exp(- \frac{\| x - x_i \|^2}{2h^2})$	$\sum_{i=1}^{n} p(x_i)(1 - \exp(- \frac{\| x - x_i \|^2}{2h^2}))$	$\sum_{i=1}^{n} \frac{p(x_i) \exp(- \frac{\|x-x_i\|^2}{2h^2})}{\sum_{i=1}^{n} p(x_i) \exp(- \frac{\|x-x_i\|^2}{2h^2})} x_i$	$\sum_{i=1}^{n} \frac{p(x_i) \exp(- \frac{\|x-x_i\|^2}{2h^2})}{\sum_{i=1}^{n} p(x_i) \exp(- \frac{\|x-x_i\|^2}{2h^2})} x_i - x$
4	$\log(1 + \exp(- \| x - x_i \|^2))$	$\sum_{i=1}^{n} p(x_i) \log(1 + \exp(- \| x - x_i \|^2))$	$\sum_{i=1}^{n} \frac{p(x_i) \frac{\exp(-\|x-x_i\|^2)}{1+\exp(-\|x-x_i\|^2)}}{\sum_{i=1}^{n} p(x_i) \frac{\exp(-\|x-x_i\|^2)}{1+\exp(-\|x-x_i\|^2)}} x_i$	$\sum_{i=1}^{n} \frac{p(x_i) \frac{\exp(-\|x-x_i\|^2)}{1+\exp(-\|x-x_i\|^2)}}{\sum_{i=1}^{n} p(x_i) \frac{\exp(-\|x-x_i\|^2)}{1+\exp(-\|x-x_i\|^2)}} x_i - x$
5	$\frac{\| x - x_i \|^2}{2h^2} \exp(- \frac{\| x - x_i \|^2}{2h^2})$	$\sum_{i=1}^{n} p(x_i) \frac{\| x - x_i \|^2}{2h^2} \exp(- \frac{\| x - x_i \|^2}{2h^2})$	$\sum_{i=1}^{n} \frac{p(x_i) \exp(- \frac{\|x-x_i\|^2}{2h^2})(1 - \frac{\|x-x_i\|^2}{2h^2})}{\sum_{i=1}^{n} p(x_i) \exp(- \frac{\|x-x_i\|^2}{2h^2})(1 - \frac{\|x-x_i\|^2}{2h^2})} x_i$	$\sum_{i=1}^{n} \frac{p(x_i) \exp(- \frac{\|x-x_i\|^2}{2h^2})(1 - \frac{\|x-x_i\|^2}{2h^2})}{\sum_{i=1}^{n} p(x_i) \exp(- \frac{\|x-x_i\|^2}{2h^2})(1 - \frac{\|x-x_i\|^2}{2h^2})} x_i - x$

Bahraini, T., Azimpour, P. and Yazdi, H.S., 2021. Modified-mean-shift-based noisy label detection for hyperspectral image classification. Computers & Geosciences, 155, p.104843.

Hyperspectral Image By Mean Shift Clustering

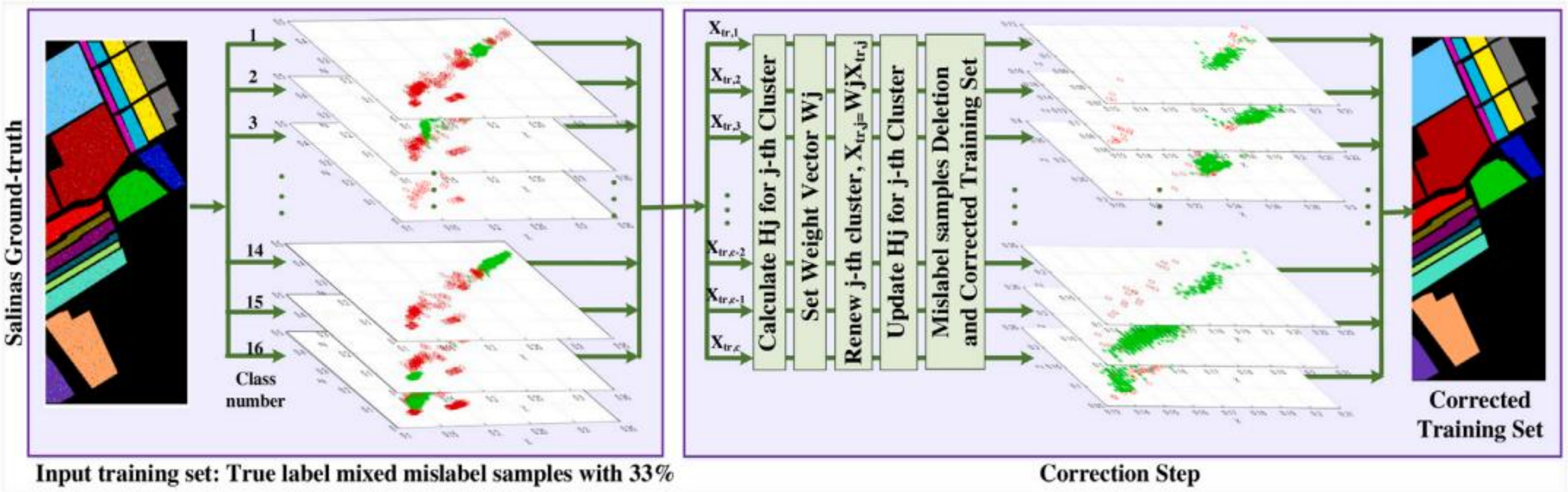


Fig. 1. Block diagram of proposed MMS method for mislabeled samples detection and correction.

Bahraini, T., Azimpour, P. and Yazdi, H.S., 2021. Modified-mean-shift-based noisy label detection for hyperspectral image classification. Computers & Geosciences, 155, p.104843.

Multi-view data fusion

- W-Step:

W-subproblem: By omitting irrelevant variables, the variable $\mathbf{w_i}$ can be obtained by solving the following problem:

$$\max_{\mathbf{w_i}} \sum_{v=1}^{V} \sum_{i=1}^{N_{tr}} \eta_v \|\mathbf{G_v^i w_i}\|_2^2 p_{vi} - \gamma Tr(\mathbf{F}^T \mathbf{L_W F}) \quad (14)$$
$$s.t. \quad \mathbf{1}^T \mathbf{w_i} = 1, \mathbf{w_i} \geqslant 0, \forall i = 1, \ldots, N_{tr}.$$

By defining $q_{ji} = -p_{ji}$, $j = 1, \ldots, V$, $i = 1, \ldots, N_{tr}$, Eq. (14) becomes

$$\min_{\mathbf{w_i}} \sum_{v=1}^{V} \sum_{i=1}^{N_{tr}} \eta_v \|\mathbf{G_v^i w_i}\|_2^2 q_{vi} + \gamma Tr(\mathbf{F}^T \mathbf{L_W F}) \quad (15)$$
$$s.t. \quad \mathbf{1}^T \mathbf{w_i} = 1, \mathbf{w_i} \geqslant 0, \forall i = 1, \ldots, N_{tr}.$$

- p-Step:

p-subproblem: We update p_{vi} and fix the other variables, and our optimization problem (13) becomes

$$\max_{p_{vi}} \sum_{v=1}^{V} \sum_{i=1}^{N_{tr}} (\eta_v \|\mathbf{G_v^i w_i}\|_2^2 p_{vi} + p_{vi} \log(-p_{vi}) - p_{vi}). \quad (18)$$

By taking the derivative of Eq. (18) with respect to p_{vi} and setting it to zero

$$p_{vi} = -\exp(-\eta_v \|\mathbf{G_v^i w_i}\|_2^2). \quad (19)$$

Zamiri, M., Bahraini, T. and Yazdi, H.S., 2021. MVDF-RSC: Multi-view data fusion via robust spectral clustering for geo-tagged image tagging. Expert Systems with Applications, 173, p.114657.

Multi-view data fusion

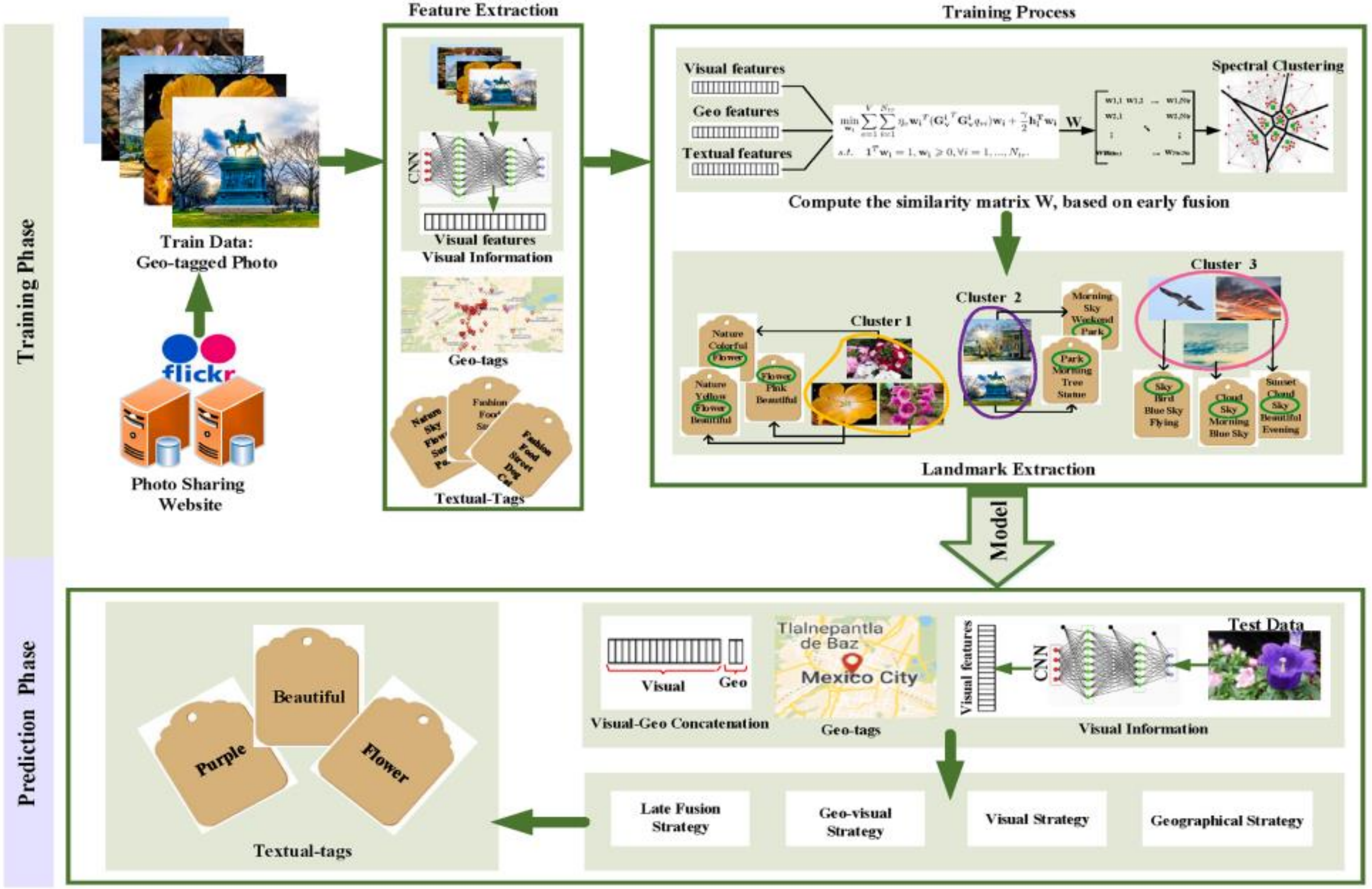


Fig. 2. Schematic illustration of the proposed image tagging method.

Zamiri, M., Bahraini, T. and Yazdi, H.S., 2021. MVDF-RSC: Multi-view data fusion via robust spectral clustering for geo-tagged image tagging. Expert Systems with Applications, 173, p.114657.

Multi-view data fusion

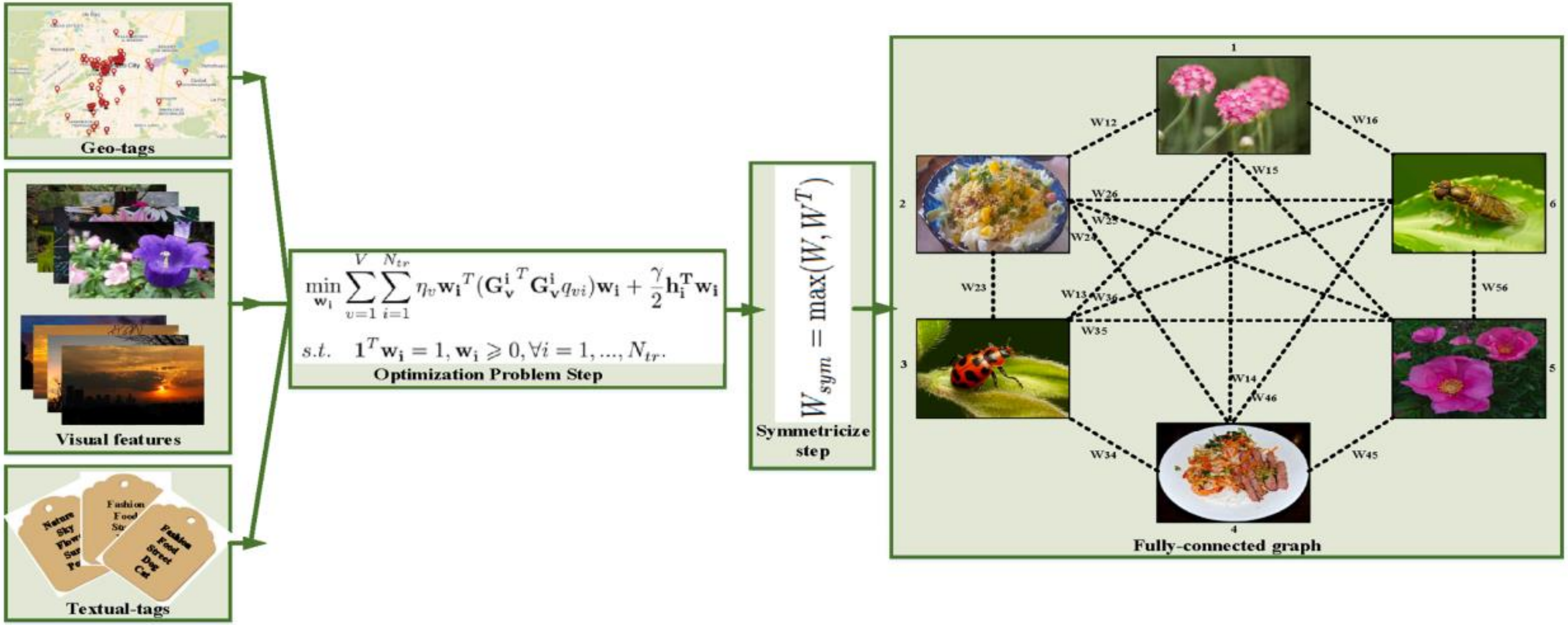


Fig. 3. An overview of the similarity matrix W_{sym} construction.

Zamiri, M., Bahraini, T. and Yazdi, H.S., 2021. MVDF-RSC: Multi-view data fusion via robust spectral clustering for geo-tagged image tagging. Expert Systems with Applications, 173, p.114657.

Multi-view data fusion

Fig. 4. Example images of Flickr (first row) and 500PX (second row) with their corresponding tags.

Zamiri, M., Bahraini, T. and Yazdi, H.S., 2021. MVDF-RSC: Multi-view data fusion via robust spectral clustering for geo-tagged image tagging. Expert Systems with Applications, 173, p.114657.

Multi-view data fusion

Fig. 11. The proposed method results for some random-selected test images for cluster "garden" of the Flickr dataset.

Zamiri, M., Bahraini, T. and Yazdi, H.S., 2021. MVDF-RSC: Multi-view data fusion via robust spectral clustering for geo-tagged image tagging. Expert Systems with Applications, 173, p.114657.

Multi-view data fusion

(a) (b)

Fig. 15. The proposed method results for Flickr dataset (a) several random-selected test images for cluster "washington" and (b) a map view of test image distribution which "washington" is assigned to them. There are peaks in locations where more images are shared.

Zamiri, M., Bahraini, T. and Yazdi, H.S., 2021. MVDF-RSC: Multi-view data fusion via robust spectral clustering for geo-tagged image tagging. Expert Systems with Applications, 173, p.114657.

Tracking

- Vehicle tracking with Kalman filter using online situation assessment
  - Vehicle tracking in the field of public transportation using Kalman filter (KF)
  - Utilizing online situation assessment (SA) inside Kalman filter is studied
  - Motion History Graph is used as online modeling of the history of the vehicle motions and is used to augment the estimation.

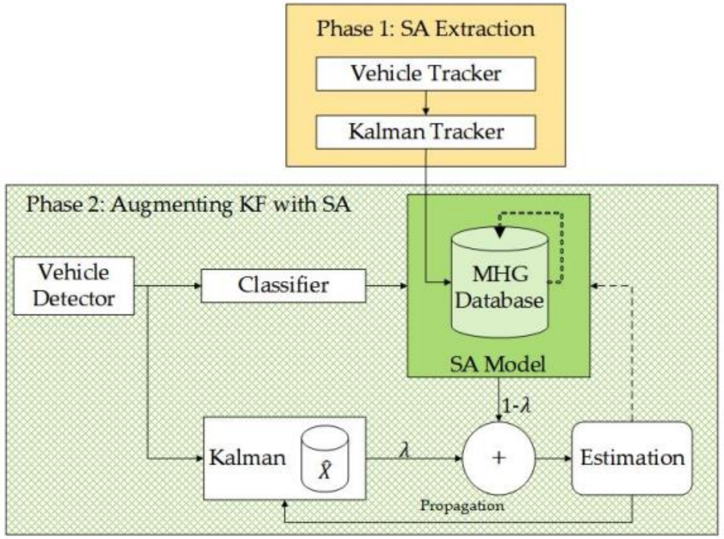


Fig. 3. Overall scheme of the proposed method

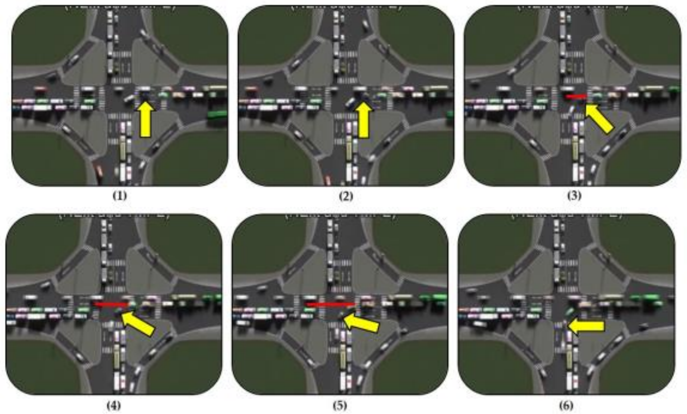


Fig. 2. How Kalman estimate with and without SA information

Khalkhali, M.B., Vahedian, A. and Yazdi, H.S., 2020. Vehicle tracking with Kalman filter using online situation assessment. Robotics and Autonomous Systems, 131, p.103596.

Tracking

Fig. 5. Samples of constructing WDG on three video sequences

Fig. 4. A sample of path extraction from video sequence

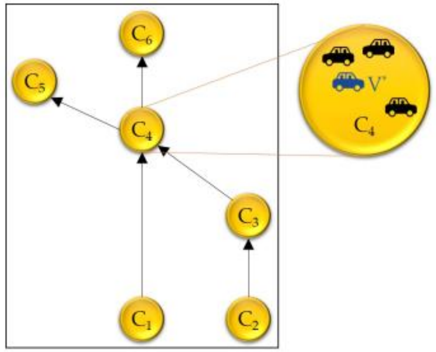


Fig. 6. A sample of clusters in WDG

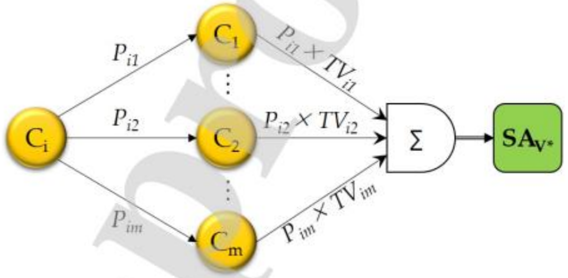


Fig. 7. How SA is calculated for each cluster center at time t

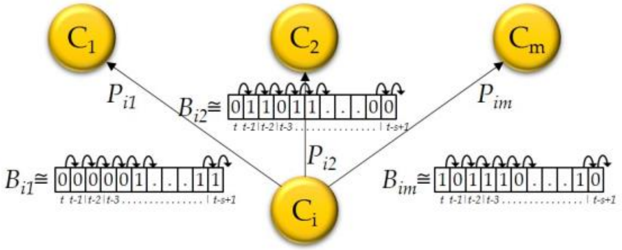


Fig. 8. The scheme of online storing of movement history for each edge at time t

Khalkhali, M.B., Vahedian, A. and Yazdi, H.S., 2020. Vehicle tracking with Kalman filter using online situation assessment. Robotics and Autonomous Systems, 131, p.103596.

Tracking

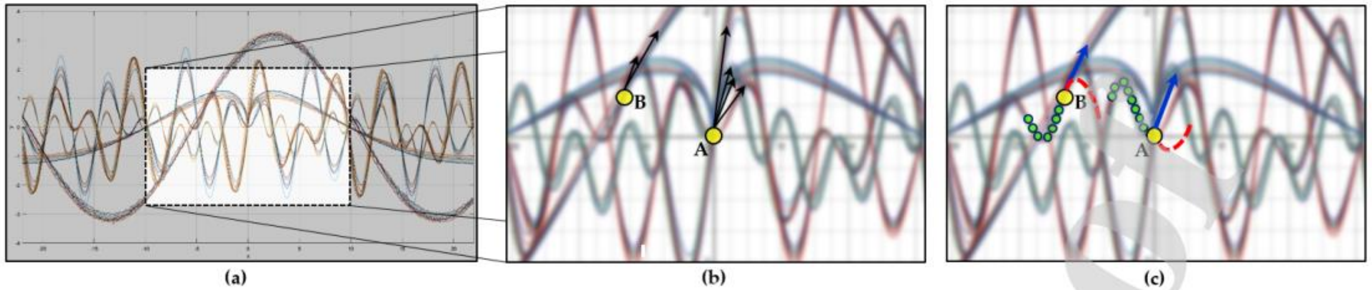


Fig. 10. Estimating sine-like signals using KF and SAKF (a) different available sine-like signals used to construct MHG (b) two sample nodes of MHG at points A and B for which outgoing edged are drawn using black arrows (c) the result of estimation using KF and SAKF: KF incorrectly estimates along with the red dashed line, while SAKF uses MHG and correctly estimates along with the blue arrow

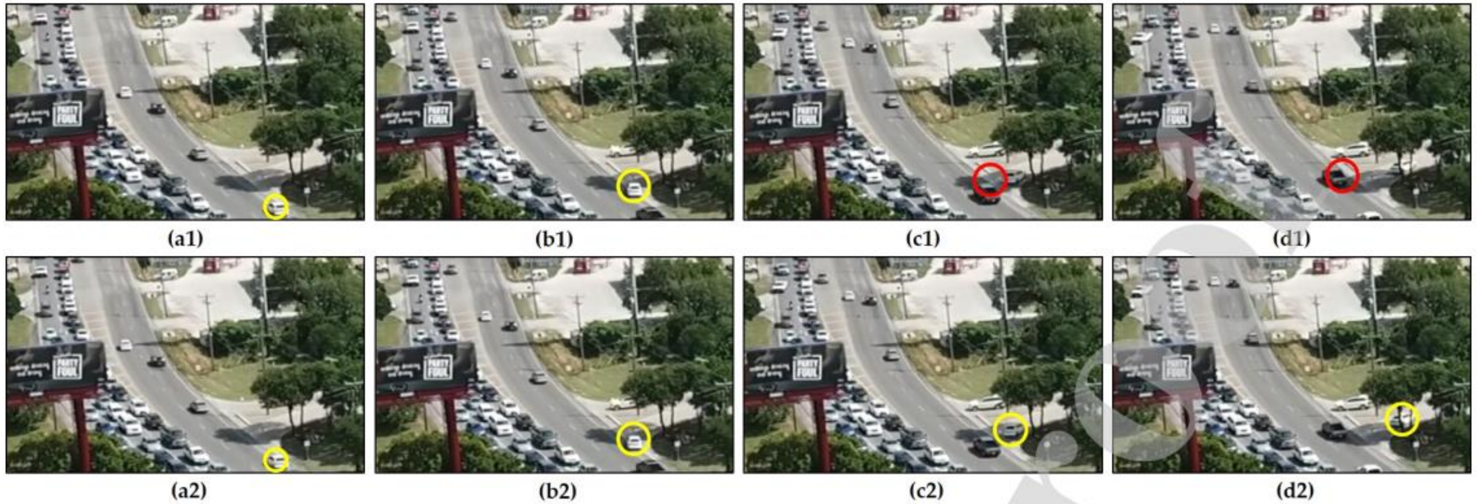


Fig. 12. Comparing KF vs. SAKF for cases where there is an option to turn for $\lambda = 0.4$

Algorithm 1 SAKF Algorithm

```
procedure SAKF()
    VS ← The video sequence with m frames
    MHG ← Construct_MHG(VS)
    for each (cluster C in MHG) do
        Calculate ΔSA(C)
    if t = 0 then
        X̂_0 ← E[X_0]
        P_0 ← E[(X_0 − E[X_0])(X_0 − E[X_0])^T] = Π_0
    while termination condition has not met do
        t = t + 1
        i ← cluster with least Euclidean distance to X̂_{t−1}
        ΔSA_t ← ΔSA(i)
        State_Estimation_Propagation() Eq. 15
        Error_Covariance_Propagation() Eq. 26
        Kalman_Gain_Matrix() Eq. 29
        State_Estimation_Update() Eq. 5
        Error_Covariance_Update() Eq. 25
```


Khalkhali, M.B., Vahedian, A. and Yazdi, H.S., 2020. Vehicle tracking with Kalman filter using online situation assessment. Robotics and Autonomous Systems, 131, p.103596.

Conclusion

- Clustering & loss function
- Human & Society based loss function

Ferdowsi University
of Mashhad

Thanks for Your Attention