Learning with Augmented Class by Exploiting Unlabeled Data AAAI 2014

Open-Category Classification by Adversarial Sample Generation IJCAI 2017

Streaming Classification with Emerging New Class by Class Matrix Sketching AAAI 2017

2018.3.15

LAC (learning with augmented class) 问题：

给定一个训练数据集 $D=\{(x_i,y_i)\}_{i=1}^{L}$， $x_i\in R^d$， $y_i\in Y=\{1,2,\ldots,K\}$。与传统分类器不同的是在测试阶段预测的样本来自一个开放数据集 $D_o=\{x_i,y_i\}_{i=1}^{\infty}$，$y_i\in Y=\{1,2,\ldots,K,K+1,\ldots,M\}$，$M>K$。由于存在训练阶段未观测到的类，目标是学习一个模型 $f(x):X\to Y'=\{1,2,\ldots,K,novel\}$，$novel$ 表示 x 属于新增类，目标是最小化如下期望风险：

$f^*=\mathrm{argmin}_{f\in\mathcal{H}}\mathbb{E}_{(x,y)\sim D_o}err(y,f(x))$，其中 $\mathcal{H}$ 是一个假设空间， err 是LAC误差：

$$err(y,f(x))=\begin{cases}I(f(x)\neq y), & y\in Y\\ I(f(x)\neq novel), & y\notin Y\end{cases}$$

LACU框架（LACU Framework）

基于最大间隔分类器的成功，我们假设一些类即使未标记的类也可以通过大间隔分类器区分。因此当区分一个seen class和其他seen classes时，未标记数据也可以帮助确定许多大间隔的分类器。然后最小化这些分类器之间的新增风险（augment risk），在这些分类器中我们选择最靠近已标记区域的那个。

$f(x) \in \mathcal{H}$ 是分类器函数，$\ell_h(f, D)$ 是训练样本上的经验损失，$\ell_u(f, D_u)$ 是未标记数据位于间隔内引起的损失，$\ell_a(f, D)$ 是新增损失（augment loss），最小化它即推动决策边界向已标记样本的方向靠近。在LACU框架中，训练分类器的目标函数是

$$\min_{f \in \mathcal{H}} \|f\|_{\mathcal{H}}^2 + C_1 \ell_h(f, D) + C_2 \ell_u(f, D_u) + C_3 \ell_a(f, D) \quad (3)$$

LACU-SVM

依据多分类中one-vs-rest的策略，在LACU-SVM中，式（3）的优化要进行 K 次，每次将一个可见类样本作为正样本（positive， $y_i = 1$ ）,其他可见类样本做为负样本（negative, $y_i = -1$ ）。$f(x) = \boldsymbol{w}^T \phi(x) + b$ 表示线性分类器.

已标记数据的hinge loss $\ell_h(f, D)$ 为：

$$\ell_h(f, D) = \sum_{i=1}^{L} \max(0, 1 - y_i f(x_i))$$

为了获得针对未标记数据的大间隔分类器，定义 $\ell_h(f, D_u)$ 为分布在在分类器间隔内的数据产生的损失之和，

$$\ell_u(f, D_u) = \sum_{i=L+1}^{L+U} \max(0, 1 - |f(x_i)|)$$

通过调节最小间隔值来最小化augment loss，从而控制分类边界的移动：

$$\ell_a(f, D) = \min_{i \in I^+} y_i f(x_i) - \min_{i \in I^-} y_i f(x_i)$$

其中， I^+ 和 I^- 分别是训练数据中正负样本的indices。

利用以上损失的式（3）训练 K 个分类器后，对一个预测实例 x ,使用如下规则对其预测：

$$\hat{y} = \operatorname{argmax}_{k=1,\ldots,K,novel} f_k(x)$$,

其中令 $f_{novel} \equiv 0$,即值为0表示预测为 $novel$ 。

求解

然而LACU-SVM里式（3）中的目标函数复杂，因此考虑可替代的目标函数：

$\min_{f\in\mathcal{H}} \|f\|_{\mathcal{H}}^2 + C_1\ell_h(f,D) + C_2\ell_u(f,D_u)$ ，并且有一个约束条件：

$$\min_{i\in I^+} y_i f(x_i) - \min_{i\in I^-} y_i \leq -\lambda$$

这里 $\lambda > 0$ 是一个参数来控制决策边界接近正样本的程度，通过消去一个最小函数上式和下式等价：

$$\min_{i\in I^+} y_i f(x_i) + \lambda \leq y_j f(x_j), \forall j \in I^-, \quad （4）$$

尽管如此，由于 $\ell_u(f,D_u)$ 中的对称hinge loss目标函数还是过于复杂。从（Collobert et al.2006）,对称hinge loss可以通过如下方式近似：

$$\max(0,1-|z|) \approx R_s(z) + R_s(-z) + constant$$

这里 $R_s(z) = min(1-s, max(0,1-z))$ 是超参数为 $s \in (-1,0]$ 的ramp loss。它可以重新写为两个hinge loss的差，也就是 $R_s(z) = H_1(z) - H_s(z)$,其中 $H_s(z) = max(0, s-z)$

$$\min_\theta J(\theta) = J_1(\theta) + J_2(\theta) \quad (5)$$

$$s.t.\ \min_{i \in I^+} y_i f(x_i) + \lambda \leq y_j f(x_j), \forall j \in I^-$$

$$\frac{\eta}{L} \sum y_i \leq \frac{1}{U} \sum_{i=L+1}^{L+U} f(x_i) \leq \frac{1}{L} \sum_{i=1}^{L} y_i \quad (6)$$

$$y_i = +1 \text{ for } L+1 \leq i \leq L+U$$

$$y_i = -1 \text{ for } L+U+1 \leq i \leq L+2U$$

$$x_{L+U+i} = x_{L+i} \text{ for } 1 \leq i \leq U$$

$$J_1(\boldsymbol{\theta}) = \|f\|_{\mathcal{H}}^2 + C_1 \ell_h(f, D) + C_2 \sum_{i=L+1}^{L+2U} H_1 y_i f(x_i)$$

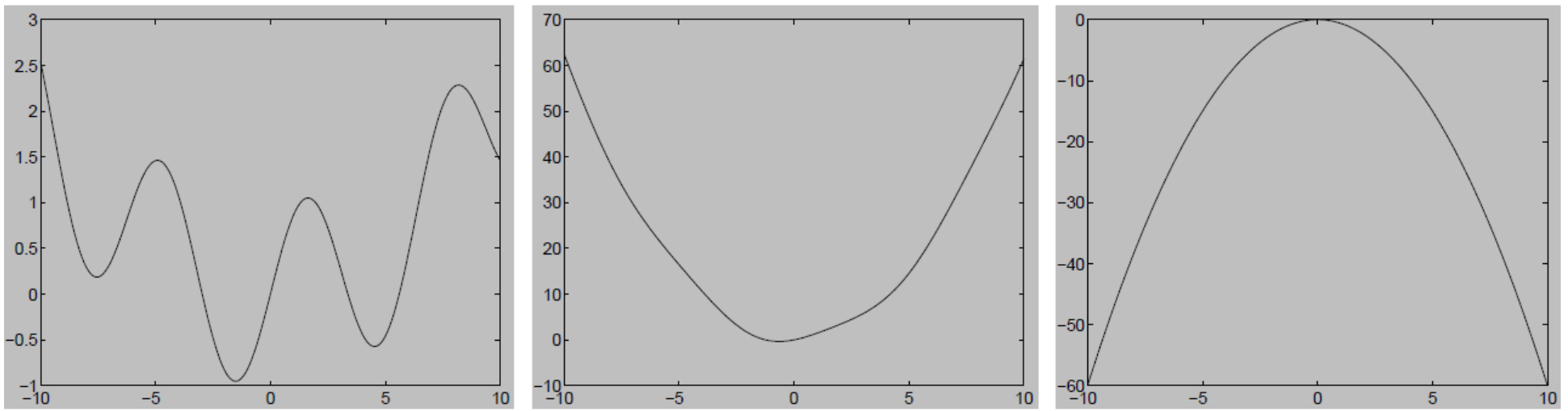
$$J_2(\boldsymbol{\theta}) = -C_2 \sum_{i=L+1}^{L+2U} H_s(y_i f(x_i))$$

CCCP (concave-convex procedure)

$$\boldsymbol{\theta}_{t+1} = \text{argmin}_{\boldsymbol{\theta}}(J_1(\boldsymbol{\theta}) + J_2'(\theta_t) \cdot \boldsymbol{\theta}), \quad (7)$$

The Concave-Convex Procedure (CCCP)（NIPS2003）

Theorem 1： 设 $E(\vec{x})$ 是一个能量函数，并且 $\dfrac{\partial^2 E(\vec{x})}{\partial \vec{x} \partial \vec{x}}$ 有界。那么我么总能将它分解成一个凸函数和一个凹函数的和。如下图：



Theorem 2： 考虑一能量函数 $E(\vec{x})$（有下界），其形式为 $E(\vec{x}) = E_{vex}(\vec{x}) + E_{cave}(\vec{x})$，其中 $E_{vex}(\vec{x})$ 和 $E_{cave}(\vec{x})$ 分别是凸函数和凹函数。那么可以由下式给出离散迭代的CCCP算法 $\vec{x}^t \longmapsto \vec{x}^{t+1}$ ：$\vec{\nabla} E_{vex}(\vec{x}^{t+1}) = -\vec{\nabla} E_{cave}(\vec{x}^t)$ 。并且CCCP可以确保能量函数 $E(\vec{x})$ 随迭代进行单调下降，因此可以收敛到最小值或者鞍点。

Theorem 3：设 $E(\vec{x}) = E_{vex}(\vec{x}) + E_{cave}(\vec{x})$ ，其中 $\vec{x}$ 满足线性约束 $\sum_i c_i^\mu x = \alpha^\mu$ $\{c_i^\mu\}, \{\alpha^\mu\}$ 是约束。那么 $\vec{x}^{t+1}$ 可以写成一系列能量函数 $E_{t+1}(\vec{x}^{t+1})$ 的更新形式：

$$E_{t+1}(\vec{x}^{t+1}) = E_{vex}(\vec{x}^{t+1}) + \sum_i x_i^{t+1} \frac{\partial E_{con}}{\partial x_i}(\vec{x}^t) + \{\sum_\mu \lambda_\mu \sum_i c_{i\mu} x_i^{t+1} - \alpha_\mu\} ,$$

其中拉格朗日乘数 $\{\lambda_\mu\}$ 施加线性约束。

$$\min_\theta J(\theta) = J_1(\theta) + J_2(\theta) \quad (5)$$

$$\boldsymbol{\theta}_{t+1} = \mathrm{argmin}_{\boldsymbol{\theta}}(J_1(\boldsymbol{\theta}) + J_2'(\theta_t) \cdot \boldsymbol{\theta}), \quad (7)$$

这个子问题的目标函数变成了一个凸项和一个线性项的和。注意在式（4）中包含一个最小函数的非凸项。为了处理这个问题，在CCCP框架中结合了迭代优化技巧。在 t+1代是用一个固定值 $V = \min_{i \in I^+} y_i f(x_i|\boldsymbol{\theta}_t)$ 来代替 $\min_{i \in I^+} y_i f(x_i|\boldsymbol{\theta}_{t+1})$ 。

$$\max_{\boldsymbol{\alpha}} \boldsymbol{\zeta}^{\boldsymbol{T}} \boldsymbol{\alpha} - \frac{1}{2} \boldsymbol{\alpha}_{\boldsymbol{T}} \boldsymbol{G} \boldsymbol{\alpha}, \quad (8)$$

$$s.t. \begin{cases} 0 \le y_i \alpha_i \le C_1, 1 \le i \le L \\ -\beta_i \le y_i \alpha_i \le C_2 - \beta_i, L+1 \le i \le L+2U \\ \alpha_{L+2U+1} \le 0, \alpha_{L+2U+2} \ge 0 \\ \alpha_i \le 0, L+2U+3 \le i \le L+2U+2+|I^-| \\ \boldsymbol{\alpha}^{\boldsymbol{T}} \mathbf{1} = 0 \end{cases}$$

Given a training dataset $D = \{(x_i, y_i)\}_{i=1}^{L}, x_i \in R^d, y_i \in Y = \{1,2, \dots, K\}$. In the test phase, we need to predict the categories of an open dataset $D_o = \{(x_i, y_i)\}_{i=1}^{\infty}, y_i \in Y_o = \{1,2, \dots, K, K+1, \dots, M\}, M > K$

The goal of OCC is to learn a model $f(x): X \to Y' = \{1,2, \dots, K, novel\}$

$$err\big(y, f(x)\big) = \begin{cases} I(f(x) \neq y), & y \in Y \\ I(f(x) \neq novel), & y \notin Y \end{cases}$$

Then the problem will be easily solved by standard supervised learning

$$f^* = \arg\min_{f \in \mathcal{H}} \mathbb{E}_{(x,y)\sim\widetilde{D}} I(y \neq f(x))$$

ASG tries to find an instance that is close to the seen class instances, but is recognized as unseen class by the discriminator.

$P_D(x; D, D^-)$ denote the probability of x to be positive by the discriminator,

For class k, denote the current generated samples as D_k^-, which is empty initially. The objective that a generated sample does not belong to the seen class is

$$\arg\min_x P_D(x; D_k, D_k^- \cup \{x\})$$

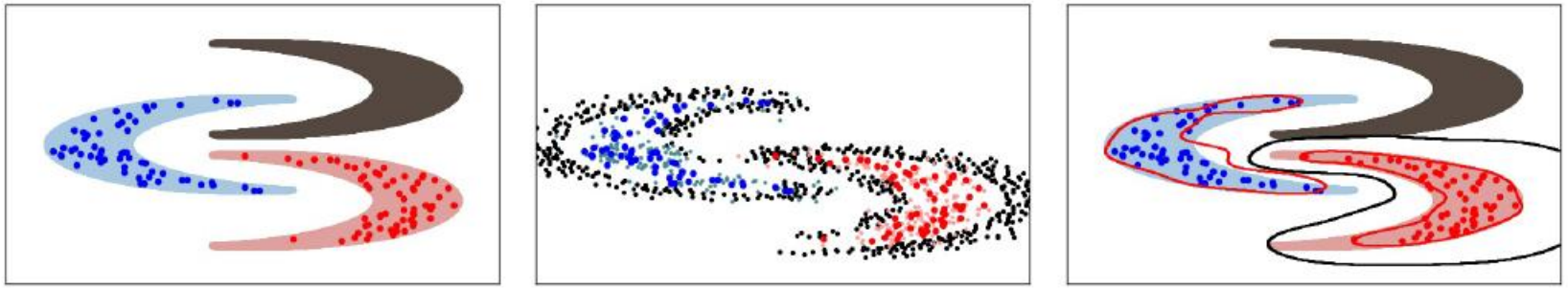
To generate boundary samples, we further require that the generated samples are close to the seen class data.

$$P_1(x, D_k) = \max\{0, \arg\min_{x' \in D_k} dist(x, x') - C_1\}$$

However, we want the generated samples to be scattered around the boundary. Therefore, we force the generated samples to be different.

$$P_2(x, D_k^-) = \max\{0, C_2 - \arg\min_{x' \in D_k^-} dist(x, x')\}$$

$$\arg\min_x \; P_D(x; D_k, D_k^- \cup \{x\}) + \lambda_1 P_1(x, D_k) + \lambda_2 P_2(x, D_k^-)$$



a) true classes and training data (b) original and generated data (c) decision boundaries

定义1：全局草图（Global Sketching(GS)）:给定训练数据集

$D = \{(x_i, y_i)\}_{i=1}^{m}, x_i \in R^d, y_i \in Y = \{1, 2, \ldots, c\}$.

全局草图为 $G \in R^{g \times d}, g \ll m$ ，矩阵 G 远远小于整个数据矩阵 $D \in R^{m \times d}$ ，但让然满足 $G^T G \approx D^T D$.

定义2：局部草图（Local Sketching(LS)）:给定一个类的数据集 $D_1, D_2, \ldots, D_c$ ，$D_i = \{(x, y)|y = i\}^{n_i}, i \in \{1, 2, \ldots, c\}$ 。局部草图为 $\mathcal{L} = \{L_1, L_2, \ldots, L_c\}, L_i \in R^{l_i \times d}, l_i \ll n_i$ 。局部草图 L_i 远小于 $D_i \in R^{n_i \times d}$ ，但仍然满足 $L_i^T L_i \approx D_i^T D_i$

Algorithm 1 Initialize Class Matrix Sketching

Input: $D \in R^{m \times d}$ - input gobal data. $D_1, D_2, \ldots, D_c \in R^{n_i \times d}$ - input local data. $G \in R^{g \times d}$, $L_i \in R^{l_i \times d}$ - all zeros matrix

Output: G and $\mathcal{L} = \{L_1, L_2, \ldots, L_c\}$

$G \leftarrow$ Construct Sketching(D,G)
for $i = 1, \ldots, c$ **do**
    $L_i \leftarrow$ Construct Sketching(D_i, L_i)
end for
    Construct Sketching(A, C)
# $A \in R^{w \times d}, C \in R^{v \times d}$
for $i = 1, \ldots, w$ **do**
    Insert A_i into a zero valued row of C
    if C has no zero valued rows **then**
        $[U, \Sigma, V] \leftarrow \text{SVD}(C)$
        $\delta \leftarrow \delta^2_{p/2}$
        $\check{\Sigma} \leftarrow \max \sqrt{(\Sigma^2 - I_p \delta, 0)}$
        $C \leftarrow \check{\Sigma} V^T$
    end if
end for
return C

给定一个测试样本 x ， $C-Mas(x)$ 产生一个类标签 $y \in \{1,2,\ldots,c,newclass\}$ 。其

中： $C-Mas(x)=\begin{cases} newclass, if \psi(x) < threhold （1） \\ j, j = max_j \phi_j(x), othwise （2） \end{cases}$

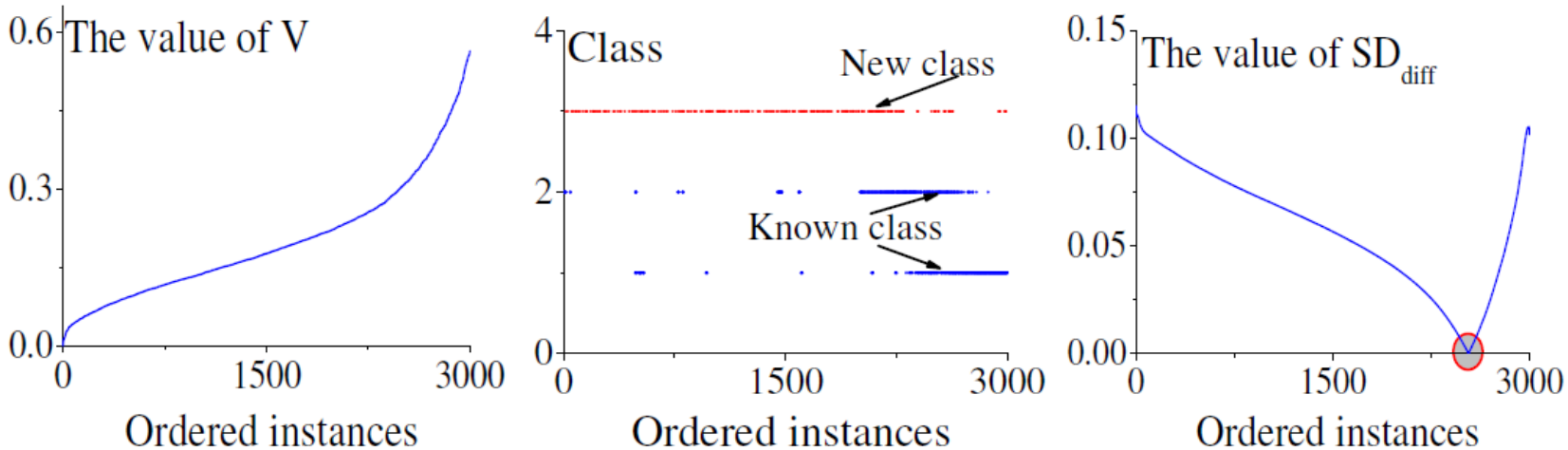
$j \in \{1,2,\ldots,c\}$, $\psi(\cdot)$ 是检测新类的函数， $\phi(\cdot)$ 是针对已知类的函数，用阈值来决定是否是新类。

设 $[\cdot]_{i,:}$ 表示一个矩阵的第 i_{th} 行。函数定义为：

$$\psi(x) = \max < x, [G]_{i,:} >, \forall i \; i \in \{1,2,\ldots,g\} （3）$$

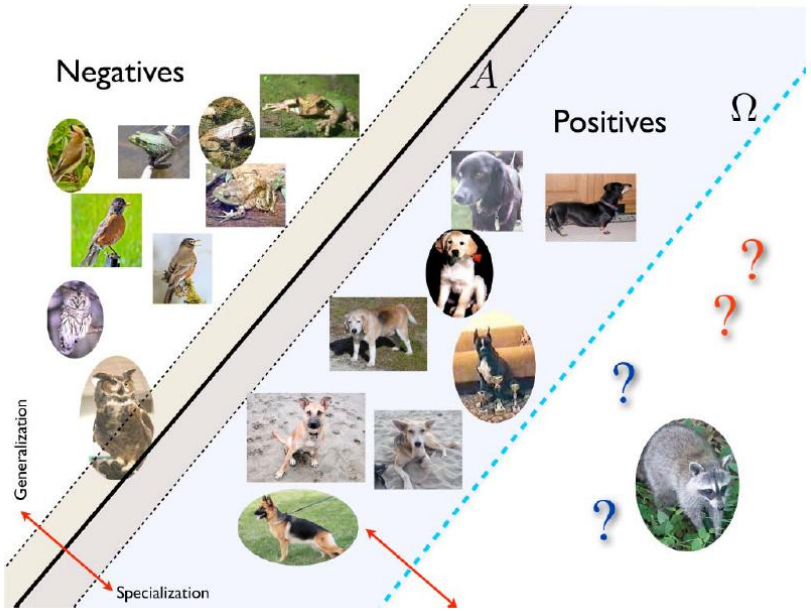
$threshold = \operatorname{argmin} Q$ （4） 或 $\frac{1}{m}\sum_{k=1}^{m} Q$ （5）　　$Q = \{\psi(x_1), \psi(x_2), .., \psi(x_m)\}$

$\hat{\tau} = \arg\min_{\tau} |\sigma(V^{left}) - \sigma(V^{right})|$ 。其中 $\sigma(\cdot)$是标准差

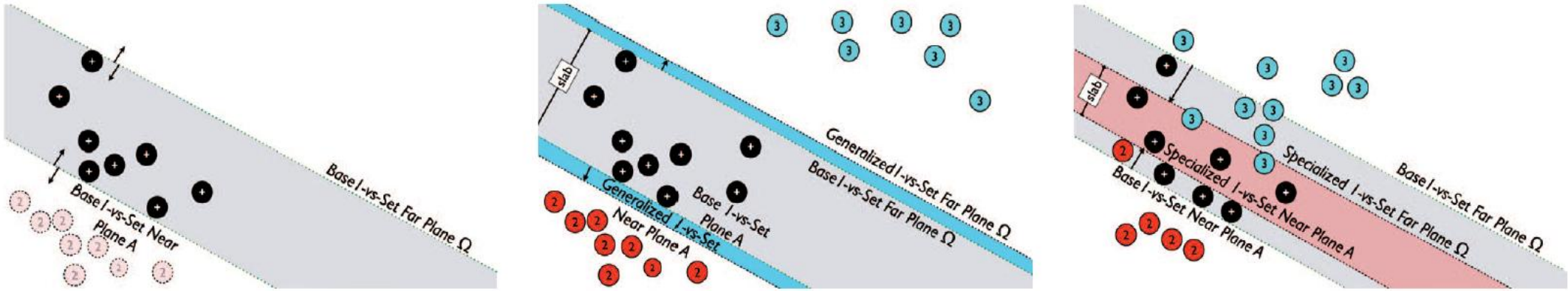


(a) V curve. (b) Class distribution. (c) SD_{diff} curve.

$$\phi_i(x) = \max < x, [L_i]_{j,:} >, \forall i, j$$
$$i \in \{1, 2, .., c\}, j \in \{1, 2, \ldots, n_j\} \quad (6)$$



Toward Open Set Recognition PAMI 2013

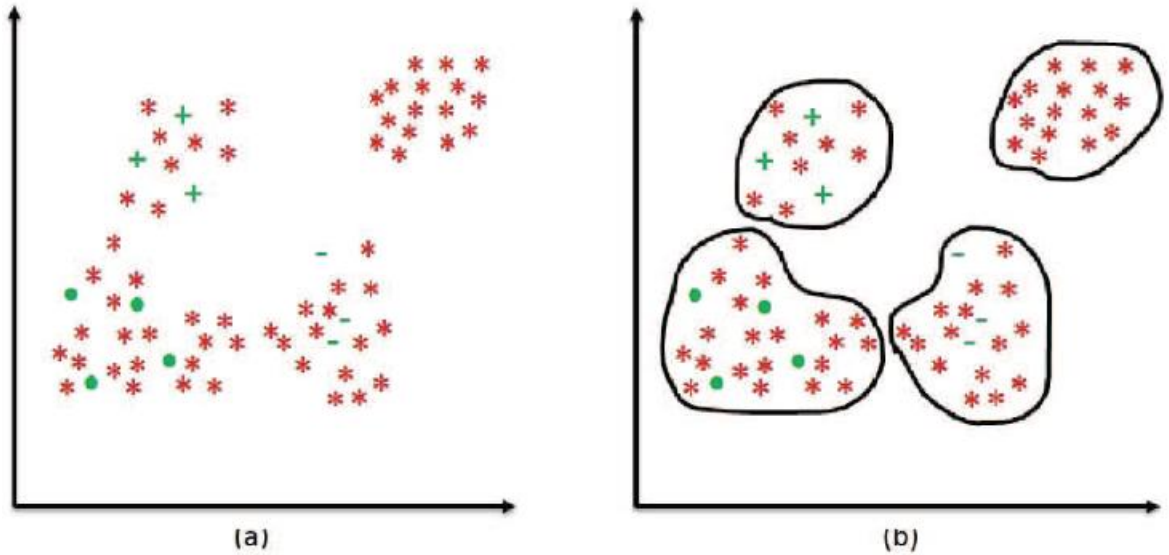


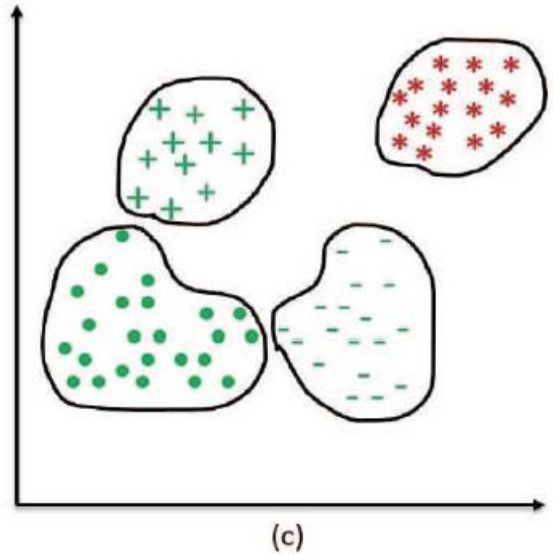
(a) Base Linear 1-vs-Set Machine (b) Generalization (c) Specialization

Fig. 4. Example of linear 1-vs-set machine showing the (a) base slab for both the 1-class and binary formulations, where the second class is only considered in the latter case, (b) the generalization, and (c) the specialization operators. Blue refers to generalization, red for specialization, and gray for the base linear 1-vs-set machine.

Detection of a New Class in a Huge Corpus of Text Documents through Semi-Supervised Learning

ICACCI 2016





Labeled Documents, Class 1

Labeled Documents, Class 2

Labeled Documents, Class 3

Unlabeled Documents

problem