指导教师：杨涛　　提交时间：3/29/2015

The task of Digital Image Processing

数字图像处理

School of Computer Science

No: 1

姓名：赵黎明

学号：2012302394

班号：10011201

基于无监督学习的孤立点自动去除方法

刘伟 华刚 约翰.史密斯

IBM T.J 沃森研究中心 斯蒂文斯理工学院

{weiliu,jsmith}@us.ibm.com ghua@stevens.edu

摘要

在计算机视觉和模式识别问题中孤立点是普遍存在的。因此，在实际收集的数据中自动消除孤立点变得越来越重要，在因特网启发性视觉应用中更是如此。本文提出一种新的学习方法，它不仅对于受污染的输入数据是健壮的，而且可以自动的发现一组数据中的孤立点。与之前的一类基于标签的监督学习方式不同，我们的方法是完全无监督的。设计中，我们优化了一个基于内核的最优目标值，通过它可以学出一个对于正常值和孤立点和分类器和软标签。一个交替优化算法可以在之后迭代的改进分类器和标签，从而可以在较少的迭代过程中实现收敛。在对于有人工添加和实际存在的孤立点的图像的研究表明，我们的方法在去除孤立点上比现阶段的所有方法效果都好，鲁棒性更高，孤立点的去除率达到 60%。

1. 介绍

最近许多研究利用从网络上获得的图像作为构建学习模型的训练数据，其中包括：学习对象的类别[8]、查询相关的视觉分类器[12]、自适应查询的图形分级[18]、语义的特定查询[28]等等。在一个典型的网络图像驱动的应用程序中，人们可以通过谷歌、必应搜索引擎或是像网络相册之类的图像分享网站抓取对于图像的文本查询（例如，对象名称或是语义概念），之后对目标物体，概念的图像集合建立模型。然而这样抓取的图像通常是带噪的，并会影响到模型学习。因此，修剪图像，也就是删除孤立点是很必要的。

对于孤立点去除和检测的现有的方法中，对于正常值形成一个标准，并且定义不符合这个标准的样本为孤立点[6,10,24,26]，或基于异常的统计或几何措施显式的隔离孤立点[7,13,17]。调查中可以发现很多方法[32]。值得注意的是许多方法都默认孤立点“数量少，且与正常值不同的”的假设，所以当孤立点的比率达到 50%的时候可能导致性能的下降。

为了更好的缓解这个问题，我们提出了一个新的方法，它可以从受污染的数据中自动的去除孤立点，以便于剩余的样本集中于一个或多个区域。我们的方法不同于以往，我们通过平等的处理正常值和孤立点来解决明显的孤立值，之后通过分类器的学习和孤立点检测规划一个简洁的学习模型用于统一特征描述。

由于联系紧密，我们的方法成为了一类学习范例[24,26]，但在输入数据的结构上是不同的。学习或是分类是从其他所有的可能对象中区分出目标数据对象。这样的一个学习问题由于消极样例的缺乏与传统监督学习方法比起来显得更加困难。由于数据获取的困难性和昂贵的人工标注，消极样例往往是不充分的。所以学习方法更受欢迎，其在图像恢复[5]、文献分

类[20]、网页分类[31]和数据流挖掘[16]等领域得到了广泛应用。

尽管如此，目前的一类学习方法例如，典型的*一类支持向量机*[24]，尽管它可以容纳一定数量的孤立点，但他并没有明确处理不确定的输入数据。另一个著名的方法*支持向量数据描述*[26]发现当把孤立点样本整合到学习方法中将提高分类器分类精度。然而，通常情况下在学习任务中我们并不知道孤立点。假如没有特定的输入配置，也没有事先的标注，我们的方法也处理这种不确定的数据混合问题。假设我们的方法处理不确定的数据混合，没有特殊的输入配置，那里既没有正确标本，也没有提前标记离群值。

我们的方法带来了新的从学习的角度来调查异常检测问题。其核心思想是定制的无监督学习机制，以应对输入数据的不确定性。其中不确定的离群值通过自导标记过程逐渐发现，然后从值得信赖的正确标本分离，培养了一大边缘单类分类。在此过程中，我们的方法不仅能有效地去除离群值，也可产生任何断言为“积极”样本的置信值。

因此，所推荐的单类学习方法可以很容易地应用到两个新兴的互联网视觉应用：网络图片标签去噪[19]和网络图片搜索重新排名[12 ， 18] 。对于标签去噪,确定的异常图像不应该采取相应的标记;对于再排序，整个图像根据我们的方法生产的置信度重新排名。由三大公共图像数据库进行的大量的实验，由我们自己人工和现实世界的离群收集新的 Web 图像数据库都证明了我们的方法的显著的性能提升，先进的异常值去除技术以及单类学习方法。

2. 相关工作

在进行异常检测和去除之前，我们做了大量的相关工作，调查了包括计算机视觉以及数据挖掘在内的很多领域。

第一类方法从几何学的视角，利用样本重建来进行异常检测。特别的是，我们可以使用由 PCA 获取的主子空间重构样本，包括 Kernel PCA, Robust PCA [4, 14, 29], 还有 Robust Kernel PCA [22, 29]，或者总结整个数据集的代表。因此异常值被确定为高残留重建样品。方法的这种风格依然遵循有关离群基本的“数和而不同”的假设。

第二类方法将异常值检测问题作为概率建模过程对待，导出了基于概率密度函数的离群度度量。这样，低概率的样品值被确定为异常值。遵从这一原则，我们探讨了巨大概率密度估计方案，包括参数估计和非参数估计，例如核密度估计（KDE）(即经典的罗森布拉特窗口方法[23])，以及更近的强劲内核密度估计（RKDE)。

第三类方法通过学习一个紧凑的数据模型来描述正常对象，使得尽可能多的正常样品内部封闭，而不是探究异常值的特质。两种常用模型有超平面和超球面。前者是由单类支持向量机(OC-SVMs)[24]提出的，而后者由支持向量数据描述 (SVDD)[26]所倡导。已被证明，OC-SVMs 和 SVDD 在采用固态内核（k(x,x)是静态不变的）时是相同的。虽然已经展示出优点，他们都需要一个由已知正常实例组成的正确的的实验集，标以“正确” ，学习超平面和超球面模型。如果实验集被损坏与相当大的比例（例如，50％)，OC-SVMs 和 SVDD 的表现都极有可能恶化，因为他们没有明确地处理过程模型的训练异常值。

让我们用一个玩具的例子来说明 OC-SVMs 和 SVDD 的异常问题。图 1(a)显示了 2D 玩具数据集合，其中的异常值来源于均匀分布的随机噪声和地面实况积极（即，常规）样品位于中心的大和密集群内。需要注意的是离群

值比例高达 50 %。采取这种损坏的数据组作为训练集，OC -SVM 被偏置到正常的点的边界附近的离群，并与判断阳性样品 RKD Einprecision 媲美，如图 1(b)(c)所示。需要注意的是，我们使用的是高斯核所以 OC- SVM 和 SVDD 输出相同的结果。相比之下，我们提出的无监督单类学习（ UOCL ）的方法（见图 1 （ d））通过在高离群水平显示出较强的健壮性达到最高的精度。

在机器学习，存在着与在本文中讨论的不确定性学习范式的一些方法，包括监督二进制类的学习与标签不确定性[3,30]，半监督单类的学习与积极的和未标记的例子[15 ，21] ，正态分布类的无监督学习[1]等等。这些都不在本文讨论范围内。

3. 半监督单分类学习

在这一部分中，我们通过提出一个可靠的无监督学习模式自动从损坏的训练数据集抹去异常值来解决目前单类学习问题。所提出的模型是建立在两个直观假设： 1）异常值从低密度样品起源，和 2）相邻的样品趋于具有一致分类。我们把这种方法称为无监督单类学习（ UOCL ），并设计了可证明收敛的算法来解决它。

3.1. 学习模型

给定一个没有标签的数据集

$$\mathcal{X} = \{x_i \in \mathbb{R}^d\}_{i=1}^n$$

$f : \mathbb{R}^d \mapsto \mathbb{R}$ 我们希望得到一个与 OC-SVMS 相似的分类函数。

$\kappa : \mathbb{R}^d \times \mathbb{R}^d \mapsto \mathbb{R}$通过借助一个核心函数

其中包括再生希尔伯特空间（RKHS）$\mathcal{H}$，这个代表定理[25]用下边的式子描述分类函数：

$$\kappa(\cdot, x_i)$$

$$f(x) = \sum_{i=1}^{n} \kappa(x, x_i)\alpha_i, \qquad (1)$$

其中 α_i 是由得到的膨胀系数。下面我们介绍一个对于输入数据 $\mathcal{X}$ 的软标签分配

$$\mathcal{Y} = \{y_i \in \{c^+, c^-\}\}_{i=1}^n$$

其中将 c^+ 作为积极样本的权值而 c^- 作为孤立点的权值。用 $y = [y_1, \cdots, y_n]^\top$ 作为$\mathcal{Y}$的向量表示。软标签的使用帮助我们处理孤立点比率较高的情况，并把处理结果作为后验。

我通过建立 UOCL 模型求得目标对象的最小化：

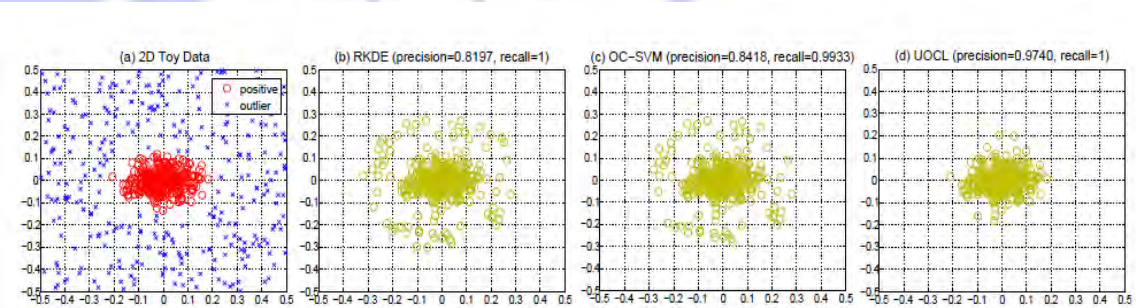

Figure 1. The outlier removal results on a 2D toy dataset where the percentage of outliers is 50%.

图 1. 在孤立点比率为 50%的 2 维玩具数据集上的去除结果

$$\min_{f\in\mathcal{H},\{y_i\}} \sum_{i=1}^{n}\left(f(x_i)-y_i\right)^2+\gamma_1\|f\|_{\mathcal{M}}^2-\frac{2\gamma_2}{n^+}\sum_{i,y_i>0}f(x_i)$$
$$\text{s.t.}\quad y_i\in\{c^+,c^-\},\ \forall i\in[1:n],$$
$$0<n^+=\left|\{i|y_i>0\}\right|<n, \tag{2}$$

其中$\gamma_1,\gamma_2>0$是控制模型的权值参数。需要注意的是，我们利用对 n+的约束 0＜n+＜n 舍弃两个极端的情况：一个积极样本没有、全是积极样本。我们设计权值(c^+,c^-)使得$\|\boldsymbol{y}\|^2$成为常量，这样做的目的是为了避免$\|\boldsymbol{y}\|^2=\sum_{i=1}^{n}y_i^2$的变化对式(2)的最优化产生影响。例如：$(1,-1)$, $(\sqrt{\frac{n}{2n^+}},-\sqrt{\frac{n}{2(n-n^+)}})$, $(\sqrt{\frac{n-n^+}{n^+}},-\sqrt{\frac{n^+}{n-n^+}})$就满足要求。有必要指出，当式(2)使用不确定的参数$\|\boldsymbol{y}\|^2$作为软标签的时候将会导致结果发散。举例来说，当使用$(\frac{1}{n^+},-\frac{1}{n-n^+})$作为软标签的时候，总会使得$n^+\approx\lfloor n/2\rfloor$。

由于在学习之前既没有积极研数据也没有孤立点标注，所以 UOCL 是一个完全的无监督学习方法，它使用平方差$(f(x_i)-y_i)^2$，而不是监督或半监督的学习方法中所使用的连接差。

式(2)中的术语$\|f\|_{\mathcal{M}}^2$是一个多维的正则矩阵，我们通过使用邻接图 G 来定义这个概念，其中邻接图中的矩阵由下式定义：

$$W_{ij}=\begin{cases}\exp\left(-\dfrac{\mathcal{D}(x_i,x_j)}{\varepsilon^2}\right), & i\in\mathcal{N}_j \text{ or } j\in\mathcal{N}_i,\\ 0, & \text{otherwise,}\end{cases} \tag{3}$$

其中$\mathcal{D}(,)$是$\mathbb{R}^d$中的距离度量。集合$\mathcal{N}_i\subset[1:n]$包含$\tilde{\mathcal{X}}$中与$\boldsymbol{x}_i$最接近的 k 个邻接的索引。之后，我们定义了一个对角矩阵 D,对角元素为$D_{ii}=\sum_{j=1}^{n}W_{ij}$，并且计算图的拉普拉斯算子矩阵$\mathbf{L}=\mathbf{D}-\mathbf{W}$ [2]之后我们可以写出如下的调整：

$$\|f\|_{\mathcal{M}}^2=\frac{1}{2}\sum_{i,j=1}^{n}\left(f(x_i)-f(x_j)\right)^2W_{ij}=f^{\top}\mathrm{L}f, \tag{4}$$

其中，向量$\boldsymbol{f}=[f(\boldsymbol{x}_1),\cdots,f(\boldsymbol{x}_n)]^{\top}$是式 1 规定的函数$f$在数据集$\tilde{\mathcal{X}}$上的实现。为了符号的简洁性，我们定义系数矢量$\boldsymbol{\alpha}=[\alpha_1,\cdots,\alpha_n]^{\top}\in\mathbb{R}^n$、内核矩阵$\mathbf{K}=[\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j)]_{1\le i,j\le n}\in\mathbb{R}^{n\times n}$、矢量内核映射$\boldsymbol{k}(\boldsymbol{x})=[\kappa(\boldsymbol{x}_1,\boldsymbol{x}),\cdots,\kappa(\boldsymbol{x}_n,\boldsymbol{x})]^{\top}$，那么目标分类函数就可以表示成为

$$f(\boldsymbol{x})=\boldsymbol{\alpha}^{\top}\boldsymbol{k}(\boldsymbol{x})\ \text{and}\ \boldsymbol{f}=\mathbf{K}\boldsymbol{\alpha}.$$

问题(2)中的表示$-\sum_{i,y_i>0}f(x_i)/n^+$在判断积极样本中有很大作用。由于缺乏准确的标签，我们考虑平均收益，而不是像 SVMs 和 OC-SVNs 一样只考虑单一收益。在积极样本上使用收益最大化策略，可以抑制由孤立点产生的偏差，从而使得大多数积极样本能够远离判定边界$f(\boldsymbol{x})=0$，提高判定的正确性。同时，为了避免无限优化，我们可以通过稳定$\|\boldsymbol{\alpha}\|=1$进一步的约

束范围$\{f(x_i) = \alpha^\top k(x_i)\}_{i=1}^n$。因此有，

$\sup\{f(x_i)|1 \le i \le n\} = \max_{1\le i\le n} \|k(x_i)\|$.

通过合并式(4)，并且忽略常数项$\|y\|^2$的情况下，我们重写问题(2)。

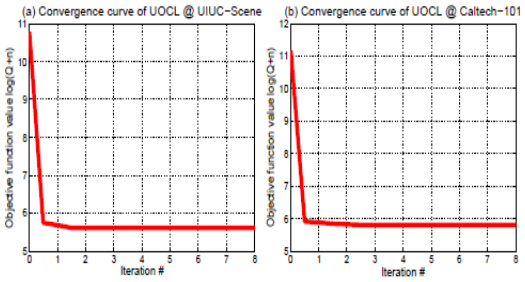

图2. UOLC的收敛性测试。

在第t次迭代时，目标函数值趋近于$\ln\left(Q(\alpha_t, \tilde{y}_t) + n\right)$。（a）UIUC-Scene的”MITcoast”类（360个样本），其中30%的异常值；（b）Caltech-101的“Faces”类（435个样本），其中30%异常值。

$$\begin{aligned} \min_{\alpha,\tilde{y}} \quad & Q(\alpha,\tilde{y}) := \alpha^\top \mathbf{K}(\mathbf{I}+\gamma_1\mathbf{L})\mathbf{K}\alpha - 2\alpha^\top \mathbf{K}\tilde{y} \\ \text{s.t.} \quad & \|\alpha\| = 1,\ \tilde{y} \in \left\{c^+ + \frac{\gamma_2}{\|\tilde{y}\|_+},\ c^-\right\}^{n\times 1}, \\ & 0 < \|\tilde{y}\|_+ < n, \end{aligned} \tag{5}$$

上式中，$\|\alpha\|$代表向量a中积极因素的个数，新的标签分配矢量$\tilde{y}$与y取相同的标志 。而问题（5）的目标函数Q是凸的，该可行解集合是不是一个凸集，使得问题（5）成为一个组合优化问题。为了突出推荐式UOCL模式的独特性，我们指出用Eq（5）公式表示的UOCL可以工作在一个自导机制下。制造更大的利润和相邻点光滑的单类分类法f。以及可以直接找出正常值和异常的软标签分配$\tilde{y}$（相当于y）。与以往的去除异常值和单类学习方法不同，我们的UOCL模型并不过分强调阳性样本或异常值。取而代之的是，它把正常和异常公平对待，使他们互相竞争，通过优化标签分配$\tilde{y}$与软标签（C+，C-）。

3.2算法

实现UOCL模型的问题（5）并不值得去解决，因为它是一个涉及连续变量α和离散变量$\tilde{y}$的混合程序。在这里，我们设计了交替优化算法与其中蕴藏着的理论基础融合并取得了良好的解决方案。

首先，我们考虑了在$\tilde{y}$一定时问题5的α-子问题：

$$\min_{\|\alpha\|=1} \quad \alpha^\top \mathbf{K}(\mathbf{I}+\gamma_1\mathbf{L})\mathbf{K}\alpha - 2(\mathbf{K}\tilde{y})^\top\alpha, \tag{6}$$

它属于具有约束的特征值问题，我们已在[9]中得到很好的研究。受已固定的$\tilde{y}$值得约束，这一子问题的最小α值为：

$$\alpha^*(\tilde{y}) = \left(\mathbf{K}(\mathbf{I}+\gamma_1\mathbf{L})\mathbf{K} - \lambda^*\mathbf{I}\right)^{-1}\mathbf{K}\tilde{y}, \tag{7}$$

其中 λ *是矩阵的最小实值特征值

$$\begin{bmatrix} \mathbf{K}(\mathbf{I}+\gamma_1\mathbf{L})\mathbf{K} & -\mathbf{I} \\ -(\mathbf{K}\tilde{y})(\mathbf{K}\tilde{y})^\top & \mathbf{K}(\mathbf{I}+\gamma_1\mathbf{L})\mathbf{K} \end{bmatrix}$$

接下来，我们处理了当α固定时，关于$\tilde{y}$的子问题，如下：

$$\begin{aligned} \max_{\tilde{y}} \quad & (\mathbf{K}\alpha)^\top\tilde{y} \\ \text{s.t.} \quad & \tilde{y} \in \left\{c^+ + \frac{\gamma_2}{\|\tilde{y}\|_+},\ c^-\right\}^{n\times 1}, \\ & 0 < \|\tilde{y}\|_+ < n. \end{aligned}$$

此离散优化问题似乎令人生畏，但确实可以在O(n log n)的时间内完成。下面的定理

针对整数m给定的一个简单的情况下，

$$m = \|\tilde{y}\|_+ \in [1, n-1].$$

定理1.
给定一个整数m（m ∈ [1, n − 1]），一个向量f（f ∈ R），该问题的一个最优解为：

$$\begin{aligned} \max_{\tilde{y}} \quad & f^\top \tilde{y} \\ \text{s.t.} \quad & \tilde{y} \in \left\{ c^+ + \frac{\gamma_2}{m},\ c^- \right\}^{n\times 1}, \\ & \|\tilde{y}\|_+ = m \end{aligned}$$

满足$\tilde{y}$ >0当且仅当fi是m最大值之一。
证明：
我们用反证法来证明这个定理。假设有这样一个最优解$\tilde{y}$，于是，我们定义她的正支持集合

$$\mathcal{C} = \{i | \tilde{y}_i > 0, i \in [1:n]\}$$

（|C|=m），C的补集为
$\mathcal{C}$ is $\overline{\mathcal{C}} = [1:n] \setminus \mathcal{C}$。如果我们的结论不成立，那么存在一个i（i ∈ C）、j（j ∈ C）。现在我们来构造另外一个可行解$\tilde{y}'$，它与另外一个新的正支持集合相对应
$\mathcal{C}' = (\mathcal{C} \setminus \{i\}) \cup \{j\}$，那么有：

$$\begin{aligned} (\tilde{y}')^\top f &= \left(c^+ + \frac{\gamma_2}{m}\right) \sum_{s\in\mathcal{C}'} f_s + c^- \sum_{s\in\overline{\mathcal{C}'}} f_s \\ &= \left(c^+ + \frac{\gamma_2}{m}\right) \left(\sum_{s\in\mathcal{C}\setminus\{i\}} f_s + f_j \right) + c^- \left(\sum_{s\in\overline{\mathcal{C}}\setminus\{j\}} f_s + f_i \right) \\ &> \left(c^+ + \frac{\gamma_2}{m}\right) \left(\sum_{s\in\mathcal{C}\setminus\{i\}} f_s + f_i \right) + c^- \left(\sum_{s\in\overline{\mathcal{C}}\setminus\{j\}} f_s + f_j \right) \\ &= \left(c^+ + \frac{\gamma_2}{m}\right) \sum_{s\in\mathcal{C}} f_s + c^- \sum_{s\in\overline{\mathcal{C}}} f_s = \tilde{y}^\top f, \end{aligned}$$

这说明，$\tilde{y}'$ 比$\tilde{y}$ 得到的目标值更大，因此$\tilde{y}'$ 不是最优解。通过矛盾我们得出结论：该定理成立。定理1表明问题9的一个最优解，可以简单地通过递减排序f ，然后在第m个已排序的元素之前切断即可。

算法 1 UOCL

输入：拉普拉斯矩阵K, L∈ Rn×n模型参数γ1， γ2 > 0, 以及软标签c+ > 0, c−< 0。
初始化：

$$\alpha_0 = \tfrac{1}{\sqrt{n}}, m_0 = \arg\max_{m\in[1:n-1]} (\mathbf{K}\alpha_0)^\top q(\mathbf{K}\alpha_0, m),$$
$$\tilde{y}_0 = q(\mathbf{K}\alpha_0, m_0), \mathbf{T} = \mathbf{K}(\mathbf{I} + \gamma_1 \mathbf{L})\mathbf{K}, t = 0;$$

重复：

$b_t := \mathbf{K}\tilde{y}_t$， λt :=$\begin{bmatrix} \mathbf{T} & -\mathbf{I} \\ -b_t b_t^\top & \mathbf{T} \end{bmatrix}$的最小特征值

$$\alpha_{t+1} := (\mathbf{T} - \lambda_t \mathbf{I})^{-1} b_t,$$
$$m_{t+1} := \arg\max_{m\in[1:n-1]} (\mathbf{K}\alpha_{t+1})^\top q(\mathbf{K}\alpha_{t+1}, m),$$
$$\tilde{y}_{t+1} := q(\mathbf{K}\alpha_{t+1}, m_{t+1}), t := t+1,$$

直至收敛。
输出：

一级分类$f^*(x) = \alpha_t^\top k(x)$ 以及基于训练集的软标签管理 $\tilde{y}^* = \tilde{y}_t$ 。

还包括 $\tilde{y}_i > 0$ 而之后$\tilde{y}_i < 0$。我们将这个最优解记为 $q(f, m)$。应注意，如果一些相同的元素出现在向量f中，可能会出现多个不同的$\tilde{y}$ 值产生相同的目标值。随后，我们回到原来的$\tilde{y}$ 子问题8，其最优解是由下式求得：

$$\tilde{y}^*(\alpha) = q(\mathbf{K}\alpha, m^*(\alpha)), \tag{10}$$

其中，

$$m^*(\alpha) = \arg\max_{m\in[1:n-1]} (\mathbf{K}\alpha)^\top q(\mathbf{K}\alpha, m). \tag{11}$$

值得一提的是，如果在确定最佳 $m^*(\alpha)$ 时出现平局，我们总是选择为其中最大的整数作为m的，以尽可能多地包含inliers，即可能正确的样本值。

到目前为止，我们已经解决了由难以直接解决的问题5派生出的两个子问题（6）和（8）。因此，我们可以制定一个交替优化算法来为问题（5）找到一个很好的解决方案。

通过借助方程（7）（9）（10），在算法一中，我们描述了这种优化算法，称其为UOCL ，并在定理2证明了其收敛。通过收敛标签分配 $\tilde{y}$ ，确定阳性标本或异常值只需通过检查 $\tilde{y}_i^* > 0$ 即可。

定理2. 该最优算法值在 α 与 $\tilde{y}$ 之间收敛。

证明：由交替优化策略，对任意t，有

$$\alpha_{t+1} = \arg\min_{\|\alpha\|=1} Q(\alpha, \tilde{y}_t)$$

$$\tilde{y}_{t+1} = \arg\min_{\tilde{y}} Q(\alpha_{t+1}, \tilde{y}),$$

由此，我们可以得出：

$$Q(\alpha_t, \tilde{y}_t) \geq Q(\alpha_{t+1}, \tilde{y}_t) \geq Q(\alpha_{t+1}, \tilde{y}_{t+1}),\ \forall t \in \mathbb{Z}.$$

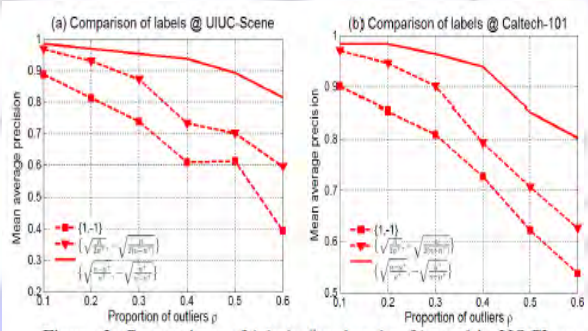

Figure 3. Comparison of labels (hard and soft) used in UOCL on the UIUC-Scene and Caltech-101 datasets.

由于 Q 从下面界的非增序列 $\{\{Q(\alpha_t, \tilde{y}_t)\}_t$ 必收敛于

$$Q^* = \lim_{t\to\infty} Q(\alpha_t, \tilde{y}_t).$$

备注：

i） UOCL 算法本质上使用核密度估计函数作为 $f_0 = \mathbf{K}1/\sqrt{n}$ 的开始，以启动它的生成。这意味着该孤立点最初寻求那些低密度样品，后来逐渐从连贯的高密度区域分离，从而对于积极样本有了更大把握。

ii）目标分类器在迭代和自我引导中训练，每次迭代 t 中，它减少噪声的标签分配 $\tilde{y}_t$，并通过对拉普拉斯图的正规化和平均收益的最大化产生平稳输出 $f_{t+1} = \mathbf{K}\alpha_{t+1}$。一个完善的标签可以通过量化 f_{t+1} 并在下次迭代中提供 f 得到。当收敛时，在 f 和 $\tilde{y}$ 中的积极数据的位置将会在输入数据集中产生一个连贯高密度子集。

iii） UOCL 算法的时间复杂度是 $O(n^3)$。事实上，UOCL 算法会在几次迭代过程中迅速收敛，这可以在式(2)中看出，对于 UOCL 来说收敛只需要三次迭代。

3.3. 讨论

既然 UOCL 算法可以在任何软标签 (c^+, c^-) 下工作，例如

$$c^+ > 0,\ c^- < 0,$$

并且 $\|y\|^2$ 是固定的，但我们希望知道当选择不同的 (c^+, c^-) 时，对 UOCL 算法将产生什么样的影响。特别是评估三种标签

$$(1,-1),\ (\sqrt{\tfrac{n}{2n^+}}, -\sqrt{\tfrac{n}{2(n-n^+)}}),$$

还有$(\sqrt{\frac{n-n^+}{n^+}},-\sqrt{\frac{n^+}{n-n^+}})$都满足$\|\boldsymbol{y}\|^2 = n$。图 3 表示在 UIUC—Scene 和 Caltech-101 数据库上使用不用的标签通过 UOCL 算法学习得到的分类器的平均准确性。图 3 的结果显示

$(\sqrt{\frac{n-n^+}{n^+}},-\sqrt{\frac{n^+}{n-n^+}})$

$(\sqrt{\frac{n}{2n^+}},-\sqrt{\frac{n}{2(n-n^+)}})$

这两个软标签要比硬标签$(1,-1)$效果好，并且$(\sqrt{\frac{n-n^+}{n^+}},-\sqrt{\frac{n^+}{n-n^+}})$效果显著的超过了其他两个，特别是孤立点比例较大的情况下。原因在于$(\sqrt{\frac{n-n^+}{n^+}},-\sqrt{\frac{n^+}{n-n^+}})$结合积极样本的数量 n+来适应标签的平衡，也就是$\sum_{i=1}^{n} y_i = 0$，所以积极样本和消极样本都以同样的方式对待。结果是自适应软标签$(\sqrt{\frac{n-n^+}{n^+}},-\sqrt{\frac{n^+}{n-n^+}})$可以使得 UOCL 算法发现并抑制孤立点的权值。需要指出的是，我们在问题(2)中的学习目标并不和 RKHS 中的正则项$\|f\|_{\mathcal{H}}^2 = \boldsymbol{\alpha}^\top \mathbf{K}\boldsymbol{\alpha}$有关，而这个正则项却在许多监督半监督核心机中出现，例如 SVNs, OC-SVMs 和 LapSVMs[2]。我们发现当把$\|f\|_{\mathcal{H}}^2$考虑进式(2)后，基本不会影响 UOCL 算法的效果。而且，我们认为 UOCL 的主要目的不是为了获得对测试数据的概括能力，而是要解决损坏的训练数据并清理孤立点。

在运行 UOCL 时，“低密度离群值的假设”已被用作初始化，其中由内核密度估计器中发现的低密度样品$f_0 = \mathbf{K1}/\sqrt{n}$被初步判定为异常值。内层数据 n+将在优化过程中改变，直到标签分配$\tilde{\boldsymbol{y}}$最终收敛。内层数据的多种模式的配置是可以处理，只要这些模式有类似的高密度。

4. 实验

我们用两个任务评估 UOCL 方法，一是离群图像去除，一是图像重排序。我们使用三个公众图象数据集，UIUC-Scene1, Caltech-1012和 INRIA[12]，同时我们整理了包含30种数据的Google-30数据库（例如，″手风琴″,″蝴蝶″,″小丑鱼″,″雄鹰″，″大象″，″壶″，″壁虎″，″帽子″，″马″，″熊猫″等）。在UIUC-Scene上，我们使用所有的15个类别。把单一品类模拟异常图像比例0.1 ≤ ρ ≤ 0.6作为从其他类别随机抽取的图像。在Caltech-101中，我们选择了11类对象每一个都包含至少100个图像，并且还模拟异常图像的比例0.1 ≤ ρ ≤ 0.6作为均匀随机采样图。在INRIA中，我们选择200条查询每个查询产生孤立值的比例0.136 ≤ ρ ≤ 0.6 ，并含有14 ？ 290幅图像。在Google-30中，15个类别产生异常的比例为0197 ≤ ρ ≤ 0.5599 ，并包含326？596幅图像。在INRIA和Google-30中，离群是现实的，这是那些不相关的图像相对于的文本查询。在所有的数据库中，内层参数和孤立点标签都是可获得的。在INRIA中，每个图像是由一个5

*1024维稀疏编码的特征向量归一化表示的。而在其它的数据集中，每个图像由21 *1024维稀疏编码特征向量表示。

我们用各种竞争方法比较UOCL ，包括五个以重建为基础的异常检测方法，PCA ，高维鲁棒PCA (HR-PCA)[29] ，内核PCA ，内核HR- PCA （ KHR-PCA ）[29] ，和稀疏建模代表选择（ SMRS ）[7] ，二个基于密度的方法，内核密度估计[23]和健壮的内核密度估计（ RKDE ）[11] 。需要注意的是OC -SVM本质上是一种监督的方法，但在本文中它是由工作在无人监督的设置下。

现在我们对于这些方法，描述孤立点测量方法。所使用的子空间方法的措施是平方重建残基和因此离群样品导致高残留。 SMRS的措施是行稀疏指数（RSI） [7]计算后的稀疏重建系数，而它只是作为临时值。而具有高行稀疏指数的临时值就被当作孤立值。对于KDE / RKDE我们采用相同的高斯核，其带宽至少通过选择方交叉验证。我们遵循[11]设置RKDE的其他参数的。对于OC- SVM / UOCL ，分类器f的输出直接表示为孤立点，即，有x使得 f(X) <0 。我们给OC-SVM/UOCL和估计 ,提供相同的高斯核 。该模型的拒绝比例参数v与OC -SVM相关，并通过最大利润的原则选择;以类似的方式，我们在UOCL中选择了模型参数γ1 ， γ2以便最大程度获得判断内围层的平均幅度。在所有的数据库中UOCL使用软标签 。

为了构造kNN图，我们定义 作为欧式距离，并固定k=6。

为了获得七种孤立点检测方法的截止阈值，我们通过确定播种测量值在孤立点上执行二进制串（两个初始种子有最大和最小值的分别）。对于每一种方法，截止阈值设置为两个集群中心的均值处。总之，所有的相比较的方法，都可以返回一个包含一个子集"判定"为积极（正常）数据的例子。除了SMRS方法不能对于所有的样例产生一个孤立点测量值，其他方法都可以根据孤立点测量值或是分类器输出对所有的样例再分级。

Table 1. **UIUC-Scene & Caltech-101** datasets: mean precision (mPre), mean recall (mRec), mean F_1 score (mF_1), mean average precision (mAP), and mean running time over the image categories of seven outlier detection methods and two one-class learning methods. All time is recorded in second. For each column, the best result is shown in boldface.

Method	UIUC-Scene (60% outliers)					Caltech-101 (60% outliers)				
	mPre	mRec	mF_1	mAP	Time	mPre	mRec	mF_1	mAP	Time
Initial	0.4011	1.0000	0.5726	–	–	0.4019	1.0000	0.5734	–	–
PCA	0.5352	0.8626	0.6587	0.6957	0.63	0.5058	0.8465	0.6321	0.6483	0.22
HR-PCA [29]	0.5336	0.8623	0.6577	0.6948	0.70	0.5221	0.8710	0.6520	0.6591	0.60
KPCA	0.5619	0.8294	0.6580	0.6122	0.54	0.5428	0.8154	0.6504	0.6436	0.25
KHR-PCA [29]	0.4684	**0.8999**	0.6147	0.5910	0.65	0.5073	**0.8825**	0.6428	0.6346	0.72
SMRS [7]	0.4536	0.8612	0.5933	–	1.91	0.5394	0.8690	0.6531	–	4.32
KDE [23]	0.5086	0.8851	0.6448	0.6892	**0.46**	0.4949	0.8579	0.6266	0.6470	**0.18**
RKDE [11]	0.5475	0.8943	0.6760	0.7306	0.47	0.5003	0.8736	0.6346	0.6570	0.19
OC-SVM [24]	0.5816	0.6209	0.5934	0.6350	2.23	0.5290	0.7155	0.6012	0.5981	4.74
UOCL (our approach)	**0.7027**	0.8822	**0.7754**	**0.8157**	1.31	**0.6795**	0.8587	**0.7483**	**0.8027**	2.28

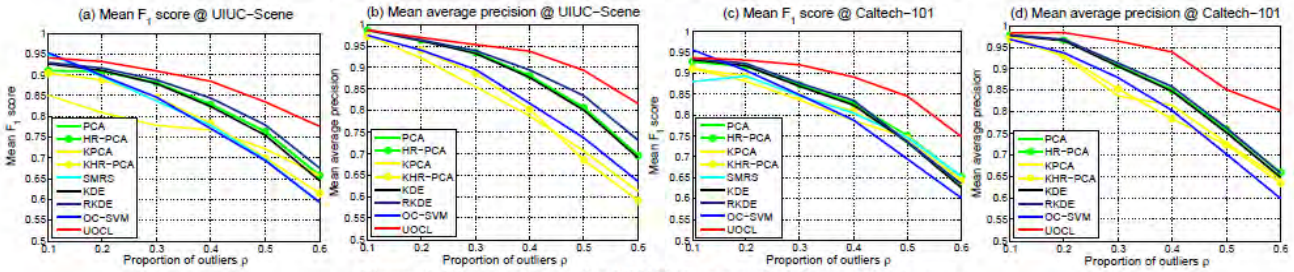

Figure 4. The results on the **UIUC-Scene** and **Caltech-101** datasets.

由于groundtruth（不知如何翻译）标签适用于所有数据集，我们可以计算精度，召回和计算所有实现去除效果的方法的得分F1，并且还计算除了SMRS方法其他的方法的重排序结果平均准确率，运行时间也被标注。结果可以在表1和表2还有图4和图5中看出。分析结果我们可以看到，UOCL方法在孤立点去除中很大程度上都能获得最高的平均精度和平均得分F1。它也始终能在重新排序方面，得到最高的平均精度和精度曲线。当孤立点的比例较大的时候UOCL的性能将远远大于其他方法。举例来说，在数据集UIUC-Scene上，在 $\rho = 0.6$的时候UOCL方法目前在mPre, mF1,和 mAP上做到21%, 15% 和 12%，比其他方法好很多。在数据集Caltech-101上mPre, mF1,和 mAP上做到25%, 15%和 22%。尽管在INRIA 和 Google-30上得到的准确性不够高，UOCL方法仍然能够达到高平均准确性、较好的平均得分F1、较好的平均精度和精度曲线。精度降低的原因可能是一些复杂查询概念的内层图像不太一致，并存在一些孤立和稀疏集群，这使得UOCL可能只捕获最密集的集群而丧失了稀疏集群。所有的这些实验结果显示：1)OC -SVM达不到的高偏离度;2）HR-PCA在一定程度上显示了稳健性，因为它可以在子空间发现并删除一些异常值;3）UOCL对离群值的图像表现出较强的健壮性，从被人工或真实世界的异常值污染的图像集中产生具有连贯性的子集。

Table 2. The results on the **INRIA** and **Google-30** datasets.

Method	INRIA					Google-30				
	mPre	mRec	mF_1	mAP	Time	mPre	mRec	mF_1	mAP	Time
Initial	0.5734	1.0000	0.7221	0.6779	–	0.7727	1.0000	0.8645	0.8476	–
PCA	0.6714	0.7183	0.6749	0.7214	**0.02**	0.8568	0.8286	0.8299	0.8940	1.21
HR-PCA	0.7033	0.7273	0.6845	0.7264	0.23	0.8637	0.8238	0.8321	0.8956	2.83
KPCA	0.6584	0.5979	0.5806	0.6988	0.03	0.8162	0.7211	0.7518	0.8488	1.34
KHR-PCA	0.6720	0.6063	0.6015	0.6883	0.35	0.8048	0.8850	0.8308	0.8648	2.12
SMRS	0.6247	0.8453	0.7055	–	0.49	0.7812	**0.9932**	0.8672	–	9.79
KDE	0.6478	0.7837	0.6916	0.7186	**0.02**	0.8380	0.8670	0.8391	0.8912	**1.12**
RKDE	0.6533	0.7807	0.6931	0.7202	0.03	0.8425	0.8680	0.8405	0.8953	1.15
OC-SVM	0.5988	**0.9505**	0.7275	0.6912	1.03	0.8144	0.9220	0.8538	0.8838	11.27
UOCL (our approach)	**0.7371**	0.8489	**0.7671**	**0.7930**	0.32	**0.9123**	0.8795	**0.8804**	**0.9119**	3.65

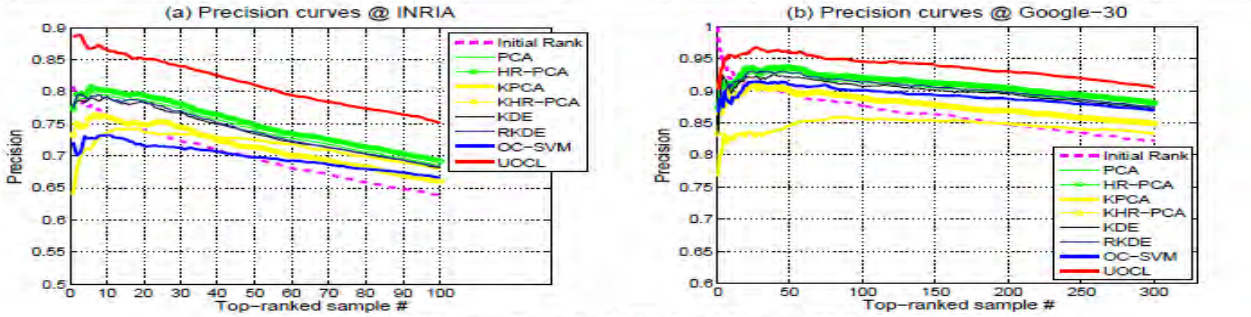

Figure 5. The results on the **INRIA** and **Google-30** datasets.

5. 结论

所提出的无监督学习方法UOCL对于受污染的输入数据是健壮的，对孤立点的抑制率能够达到60%。在四个数据集上的孤立点去除和图像重排序表明UOCL算法大大优于其余方法。UOCL的成功源于三个主要因素：1)自主学习机共同维护了了一个大容量的分类器，和内层数据、孤立点的标签。2）自适应均衡软标签被用于处理高偏离度3）交替优化算法实现了快速收敛。

致谢

刘伟博士是由约瑟夫雷维吾纪念博士后奖学金提供部分支持。华刚博士是由美国国家科学基金会提供部分支持、格兰特IIS 1350763 、中国国家自然科学基金会资助61228303 ，从华

刚的的启动资来自
史蒂文斯理工学院。

参考文献

[1] K. Balasubramanian, P. Donmez, and G. Lebanon. Unsupervised supervised learning ii: Margin-based classification without labels. JMLR, 12:3119 – 3145, 2011.

[2] M. Belkin, P. Niyogi, and V. Sindhwani. Manifold regularization: a geometric framework for learning from labeled and unlabeled examples. *JMLR*, 7:2399 – 2434, 2006.

[3] J. Bi and T. Zhang. Support vector classification with input data uncertainty. In *NIPS 17*, 2004.

[4] E. J. Cand`es, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *Journal of the ACM*, 58(3):Article 11, 2011.

[5] Y. Chen, X. S. Zhou, and T. S. Huang. One-class svm for learning in image retrieval. In *Proc. ICIP*, 2001.

[6] K. Crammer and G. Chechik. A needle in a haystack: Local one-class optimization. In *Proc. ICML*, 2004.

[7] E. Elhamifar, G. Sapiro, and R. Vidal. See all by looking at a few: Sparse modeling for finding representative objects. In *Proc. CVPR*, 2012.

[8] R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman. Learning object categories from internet image searches. *Proceedings of the IEEE*, 98(8):1453 – 1466, 2010.

[9] W. Gander, G. H. Golub, and U. von Matt. A constrained eigenvalue problem. *Linear Algebra and its Applications*, 114/115:815 – 839, 1989.

[10] G. Gupta and J. Ghosh. Robust one-class clustering using hybrid global and local search. In *Proc. ICML*, 2005.

[11] J. Kim and C. D. Scott. Robust kernel density estimation. *JMLR*, 13:2529 – 2565, 2012.

[12] J. Krapac, M. Allan, J. Verbeek, and F. Jurie. Improving web image search results using query-relative classifiers. In *Proc. CVPR*, 2010.

[13] H.-P. Kriegel, M. Schubert, and A. Zimek. Angle-based outlier detection in high-dimensional data. In *Proc. KDD*, 2008.

[14] F. D. la Torre and M. J. Black. A framework for robust subspace learning. *IJCV*, 54(1/2/3):117 – 142, 2003.

[15] W. S. Lee and B. Liu. Learning with positive and unlabeled examples using weighted logistic regression. In *Proc. ICML*, 2003.

[16] B. Liu, Y. Xiao, L. Cao, and P. S. Yu. One-class-based uncertain data stream learning. In *Proc. SIAM International Conference on Data Mining*, 2011.

[17] F. T. Liu, K. M. Ting, and Z.-H. Zhou. Isolation forest. In *Proc. International Conference on Data Mining*, 2008.

[18] W. Liu, Y.-G. Jiang, J. Luo, and S.-F. Chang. Noise resistant graph ranking for improved web image search. In *Proc. CVPR*, 2011.

[19] W. Liu, J.Wang, and S.-F. Chang. Robust and scalable graph-based semisupervised learning. *Proceedings of the IEEE*, 100(9):2624 – 2638, 2012.

[20] L. M. Manevitz and M. Yousef. One-class svms for document classification. *JMLR*, 2:139 – 154, 2001.

[21] J. Mu˜noz-Mar′ı, F. Bovolo, L. G′omez-Chova, L. Bruzzone, and G. Camps-Valls. Semisupervised one-class support vector machines for classification of remote sensing data. *IEEE Transactions on Geoscience and Remote Sensing*, 48(8):3188 – 3197, 2010.

[22] M. H. Nguyen and F. D. la Torre. Robust kernel principal component analysis. In *NIPS 21*, 2008.

[23] E. Parzen. On estimation of a probability

density function and mode. *Annals of Mathematical Statistics*, 33(3):1065－1076, 1962.

[24] B. Sch¨olkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson. Estimating the support of a high-dimensional distribution. *Neural Computation*, 13(7):1443－1471, 2001.

[25] B. Sch¨olkopf and A. J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, 2002.

[26] D. M. J. Tax and R. P. W. Duin. Support vector data description. *Machine Learning*, 54(1):45－66, 2004.

[27] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong. Locality-constrained linear coding for image classification. In *Proc. CVPR*, 2010.

[28] X. Wang, S. Qiu, K. Liu, and X. Tang. Web image re-ranking using queryspecific semantic signatures. *TPAMI*, 2014.

[29] H. Xu, C. Caramanis, and S. Mannor. Outlier-robust pca: The high-dimensional case. *IEEE Transactions on Information Theory*, 59(1):546－572, 2013.

[30] L. Xu, K. Crammer, and D. Schuurmans. Robust support vector machine training via convex outlier ablation. In *Proc. AAAI*, 2006.

[31] H. Yu, J. Han, and K. C. C. Chang. Pebl: web page classification without negative examples. *IEEE Transactions on Knowledge and Data Engineering*, 16(1):70－81, 2004.

[32] A. Zimek, E. Schubert, and H.-P. Kriegel. A survey on unsupervised outlier detection in high-dimensional numerical