Journal of Computer Applications
计算机应用,2013,33(2):326-328
ISSN 1001-9081
CODEN JYIIDU
2013-02-01
http://www.joca.cn

文章编号:1001-9081(2013)02-0326-03
doi:10.3724/SP.J.1087.2013.00326

# 基于流形主动学习的遥感图像分类算法

刘　康*,钱　旭,王自强
(中国矿业大学(北京) 机电与信息工程学院,北京 100083)
(*通信作者电子邮箱 liukang1112@gmail.com)

**摘　要**:为了高效地解决遥感图像分类问题,提出一种基于流形学习和支持向量机(SVM)的图像分类算法。在初始阶段,该算法首先利用初始训练集训练SVM,并且使用SVM找出离分类界面最近的样本;然后在所选样本中利用拉普拉斯图构建样本空间的流形结构,选出最具有代表性的样本加入训练集;最后利用高光谱图像进行实验进行验证。通过与现有的主动学习算法进行比较,结果表明该算法获得了更高的分类准确率。

**关键词**:主动学习;流形学习;拉普拉斯图;数据挖掘;机器学习
**中图分类号**:TP181　　**文献标志码**:A

# Remote sensing image classification based on active learning with manifold structure

LIU Kang*, QIAN Xu, WANG Ziqiang
(*School of Mechanical Electronic and Information Engineering, China University of Mining and Technology (Beijing), Beijing* 100083, *China*)

**Abstract:** To efficiently solve remote sensing image classification problem, a new classification algorithm based on manifold structure and Support Vector Machine (SVM) was proposed. Firstly, the proposed algorithm trained the SVM with initial training set and found the samples close to the decision hyperplane, then built the manifold structure of the samples by using Laplacian graph of the selected samples. The manifold structure was applied to find the representative samples for the classifier. The experimental evaluations were conducted on the hyperspectral images, and the effectiveness of the proposed algorithm was evaluated by comparing it with other active learning techniques exiting in the literature. The experimental results on data set confirm that the algorithm has higher classification accuracy.

**Key words:** active learning; manifold learning; Laplacian graph; data mining; machine learning



## 0　引言

监督学习模型,如支持向量机(Support Vector Machine, SVM)[1]、神经网络[2]等广泛应用于分类问题[3]。所有分类模型都需使用标记样本训练,并且分类模型的分类效果依赖于标记样本的质量。因此,训练样本需完整地表示所含类别的统计属性。然而,获取训练样本不仅费时、费力,而且训练集包含大量的冗余样本。为了尽可能地减小训练集及标注成本,在机器学习领域中提出了主动学习方法以优化分类模型。

主动学习算法可以由以下五个组件进行建模[4]:

$A = (C, L, S, Q, U)$

其中:$C$为一个或一组分类器;$L$为一组已标注的训练样本集;$Q$为查询函数,用于在未标注的样本中查询信息量大的样本;$U$为整个未标注样本集;$S$为督导者,可以对未标注样本进行标注。主动学习算法主要分为两阶段:第一阶段为初始化阶段,随机从未标注样本中选取小部分,由督导者标注,作为训练集建立初始分类器模型;第二阶段为循环查询阶段,$S$从未标注样本集$U$中,按照某种查询标准$Q$,选取一定的未标注样本进行标注,并加到训练样本集$L$中,重新训练分类器,直至达到训练停止标准为止。

主动学习算法是一个迭代的过程,分类器使用迭代时反馈的样本进行训练,不断提升分类效率。主动学习技术也应用于遥感图像分类领域[5-8],但这些算法仅仅考虑样本距离分类决策边界的距离,即将不确定性的样本用于训练分类器,而忽略所选样本多样性的问题。基于上述考虑,本文提出基于流形结构的主动学习(Active Learning Based on Manifold, ALBM)算法,在初始阶段,该算法首先利用初始训练集训练SVM,并且使用SVM找出离分类界面最近的样本;然后在所选样本中利用拉普拉斯图算子构建样本空间的流形结构,选出最具有代表性的样本加入训练集。通过与现有的主动学习算法比较,实验结果表明该算法获得了更高的分类准确率。

## 1　基于流形主动学习图像分类

### 1.1　流形学习

设$\boldsymbol{Y}$是包含在欧氏空间$\mathbf{R}^d$中的$d$维定义域,且设从$d$维欧氏空间$\mathbf{R}^d$到$D$维欧氏空间$\mathbf{R}^D$存在一个光滑的嵌入映射为:

$$f: \boldsymbol{Y} \subset \mathbf{R}^d \rightarrow \mathbf{R}^D;\ D > d \tag{1}$$

若给定有未知嵌入映射$f$生成的观测数据集$\{\boldsymbol{x}_i = f(y_i)\} \subset \mathbf{R}^D$,则流形学习的目标是从观测数据集$\{\boldsymbol{x}_i\}$重构嵌入映射和对应的低维坐标$\{y_i\}$。

上述流形学习的定义可知,只知道高维观测数据集$X = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n\}$,隐含假定观测数据集$X$位于或近似位于一个嵌入在高维欧氏空间$\mathbf{R}^D$中的内在低维流形$M$上。

在流形学习过程中,样本内在流形可以使用谱图的结构形式表示。在本文中,使用拉普拉斯图来表示样本的流形结构。设给定$l$个标记样本集$(x_1, y_1), (x_2, y_2), \cdots, (x_l, y_l)$,$u$个未标记的样本$x_{l+1}, x_{l+2}, \cdots, x_{l+u}$,通常$l \ll u$。假设$L, U$分别指

**收稿日期**:2012-08-02;**修回日期**:2012-10-16。　**基金项目**:国家自然科学基金资助项目(70701013);中国博士后科学基金资助项目(2011M500035);高等学校博士学科点专项科研基金资助项目(20110023110002)。

**作者简介**:刘康(1987-),男,湖北黄冈人,博士研究生,主要研究方向:机器学习、模式识别;　钱旭(1962-),男,江苏无锡人,教授,博士生导师,博士,主要研究方向:机器学习、信息融合;　王自强(1973-),男,河南郑州人,博士研究生,主要研究方向:机器学习、模式识别。

代标记样本集和未标记样本集,则 $n = l + u$ 表示全部的样本。假设样本标签是二元的:$y_L \in \{0,1\}$。使用 $G = (V,E)$ 表示给定 $n$ 个样本的连接图,其中,标记样本点需对应于其标签。此外,需要构造连接图边的权重矩阵 $\boldsymbol{W}_{n\times n}$。假设 $\boldsymbol{x} \in \mathbf{R}^m$,矩阵 $\boldsymbol{W}$ 定义如下:

$$w_{ij} = \begin{cases} \exp\left(-\frac{1}{\sigma^2}\sum_{d=1}^{m}(x_{id}-x_{jd})^2\right), & \boldsymbol{x}_i、\boldsymbol{x}_j \text{ 近邻} \\ 0, & \text{其他} \end{cases} \tag{2}$$

其中 $w_{ij}$ 也可以使用其他的赋值方式。因此,在欧氏空间内,给相邻的两个样本赋予较大的权重值。设 $\boldsymbol{M}$ 表示对角矩阵,则 $m_{ii} = \sum_j w_{ij}$。那么,图的拉普拉斯算子如下式所示:

$$\boldsymbol{N} = \boldsymbol{M} - \boldsymbol{W} \tag{3}$$

在本文中,$\boldsymbol{N}$ 反映了样本的内在低维流形,并且拉普拉斯图满足流形结构中平滑性的要求。

### 1.2　支持向量机

设训练样本集 $\{\boldsymbol{x}_i, y_i\}_{i=1}^{l}$,样本 $\boldsymbol{x}_i \in \mathbf{R}^n$,$y_i \in \{-1,1\}$ 为标签。SVM 通过求解式(4)找到一个具有最大间隔的超平面:

$$\min_{\boldsymbol{w},b,\xi}\left(\frac{1}{2}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{w} + C\sum_{i=1}^{l}\xi_i\right) \tag{4}$$

$$\text{s.t.}\quad y_i[(\boldsymbol{w}\cdot\boldsymbol{x}+b)+\xi_i-1] \geqslant 0,\ \xi_i \geqslant 0$$

其中:$C$ 为控制误差的惩罚常数,$\xi_i$ 为非负松弛变量。

利用 Lagrange 乘子法,将式(4)转化成对偶形式:

$$\max S(\alpha) = \sum_{i=1}^{l}\alpha_i - \frac{1}{2}\sum_{i=1}^{l}\alpha_i\alpha_j y_i y_j(\boldsymbol{x}_i\cdot\boldsymbol{x}_j) \tag{5}$$

$$\text{s.t.}\quad \sum_{i=1}^{l}\alpha_i y_i = 0, \alpha_i \in [0,C],\ i = 1,2,\cdots,l$$

对于任何未知的类别的样本,可使用下式判决函数决定其所属类别:

$$f(x) = \operatorname{sgn}\left[\sum_{i=1}^{l}\alpha_i y_i(\boldsymbol{x}_i\cdot\boldsymbol{x}_j)+b\right] \tag{6}$$

对于非线性 SVM,通过利用非线性映射 $\varphi$ 把输入空间映射到高维特征空间,于是核函数 $K(\boldsymbol{x}_i,\boldsymbol{x}_j) = (\varphi(\boldsymbol{x}_i),\varphi(\boldsymbol{x}_j))$ 可在特征空间计算,而无需知道映射 $\varphi$ 的具体形式。使用核函数替代线性 SVM 中的点积形式。

最初的支持向量机用于解决两分类问题,不能直接用于多分类,当前已有许多算法将支持向量机推广到多分类问题,现在的主要方法有:1) one-versus-rest 算法,该算法依次使用一个两类支持向量机分类器将每一类与其他所有类别区分开来,分类时将未知样本分类为具有最大分类函数值的那一类;2) one-versus-one 算法,该算法在每两类间训练一个分类器,因此当对一个未知样本进行分类时,每个分类器都对其类别进行判断,并为相应的类别投票,最后得票最多的类别即作为该未知样本的类别[9-10]。本文采用 one-versus-one 算法进行分类。

### 1.3　基于流形主动学习算法

设从图 $G$ 中找到一个实值函数 $f: V \mapsto \mathbf{R}$,并且函数在图中已标记的点被限制为 $f(\boldsymbol{x}_i) = f_l(\boldsymbol{x}_i) = y_i (i = 1,2,\cdots,l)$。在流形学习中,所有的样本点基本上落在一个低维的流形上面,沿着流形相近的点之间的类别相同,即流形假设[11-14]。基于上述原理,可定义一个二次能量函数,具体形式如下:

$$E(f) = \frac{1}{2}\sum_{i,j} w_{ij}(f(\boldsymbol{x}_i)-f(\boldsymbol{x}_j))^2 \tag{7}$$

为了表示函数 $f$ 所能表示的概率分布的形式,本文使用高斯域 $p_\lambda(f) = \exp\{-\lambda E(f)\}/S_\lambda$。其中:$\lambda$ 是逆温度参数,$S_\lambda$ 为分区函数,$S_\lambda = \int_{f|L=f_l}\exp\{-\lambda E(f)\}\mathrm{d}f$。因此,通过最小化能量函数可以求出函数 $f$ 的最优解:

$$f = \arg\min_{f|L=f_l} E(f) \tag{8}$$

由谱图理论可知,函数 $f$ 的最优解满足调和函数的性质,即:在未标记样本中满足 $\Delta f = 0$,其中 $\Delta$ 表示图拉普拉斯算子(即式(3)中的 $\boldsymbol{N}$)。为了将调和函数应用于矩阵计算,将权重矩阵和对角矩阵分别划分成 4 块:

$$\boldsymbol{W} = \begin{pmatrix} \boldsymbol{W}_{ll} & \boldsymbol{W}_{lu} \\ \boldsymbol{W}_{ul} & \boldsymbol{W}_{uu} \end{pmatrix} \tag{9}$$

同时将函数 $f$ 的解分成两部分,形象表示成 $f = \begin{bmatrix} f_l \\ f_u \end{bmatrix}$。其中:对标记样本而言,$f_l$ 表示函数解 $f$ 的值,它是已知的,即标记样本的标签值;对未标记样本而言,$f_u$ 表示函数解 $f$ 的值。因此,根据调和函数性质,$f_u$ 可表示为

$$f_u = (\boldsymbol{D}_{uu} - \boldsymbol{W}_{uu})^{-1}\boldsymbol{W}_{ul}f_l \tag{10}$$

通常,此函数需要融合一个只使用标记样本集训练得到额外的分类器。在本文提出的算法中,使用 SVM 分类器,用 $g_u$ 表示。

在原始图中,对于每一个未标记的样本点 $i$ 而言,寻找锚点 $j$(即:锚定一个标记的样本点)并同时给出其标签值 $g_j$,设 $\eta$ 表示从样本点 $i$ 到锚点 $j$ 的转移概率,$1-\eta$ 表示样本点 $i$ 到所有其他点(除了点 $j$)的转移概率。当使用调和函数最小化能量函数时,SVM 分类器引入了“标记成本”。假设 $\boldsymbol{P} = \boldsymbol{M}^{-1}\boldsymbol{W}$,由随机游走思想可得函数

$$f_u = (I - (1-\eta)P_{uu})^{-1}((1-\eta)P_{ul}f_l + \eta g_u) \tag{11}$$

理论上而言,最小化问题可以看作选择最优的训练样本集。因此,可定义一个风险函数,其包括两个部分:SVM 和调和函数。通过最小化风险函数,可选出最优的子集。优化函数形式如下:

$$x_k^* = \arg\min_{x_k \subset D_u^i}\{(1-\eta)R(f^{x_k}) + \eta R(g^{x_k})\} \tag{12}$$

为了最小化 SVM 的风险值,目前最常用的方法是选择最靠近决策边界的样本,尽可能地减小解空间,从而达到减少风险值的效果。对于调和函数 $f_u$ 而言,可通过式(13)求出最优解:

$$f_u^{+(x_k,y_k)} = f_u + (y_k - f_k)\frac{(\Delta_{uu}^{-1})_{\cdot k}}{(\Delta_{uu}^{-1})_{kk}} \tag{13}$$

其中:$(\Delta_{uu}^{-1})_{\cdot k}$ 表示未标记样本逆拉普拉斯矩阵第 $k$ 列;$(\Delta_{uu}^{-1})_{kk}$ 表示逆拉普拉斯矩阵的第 $k$ 个对角元素。

由上式可见,通过迭代过程,算法可以逐步地选出最优的样本加入训练集,减小分类风险值,优化分类器,提高分类效率。

## 2　实验结果

为了评估本文算法(ALBM)的有效性,使用已有的主动学习算法与之比较:随机抽样(Random sampling,RANDOM)算法、边缘抽样(Margin Sampling,MS)算法和基于多层次不确定性抽样(MultiClass-Level Uncertainty,MCLU)算法。在 MS 算法中,每次迭代都选择距离决策界面最近的样本加入训练集;而 MCLU 考虑距离分类界面最远的两个点的差值,并选择差值最小的样本加入训练集。

首先使用 UCI 数据库中的 wine 数据集验证各主动学习算法的准确率如图 1 所示。从中可以看出本文的 ALBM 算法的分类准确率高于其他算法;换言之,谱图能够很好地表示该

数据集的流形。

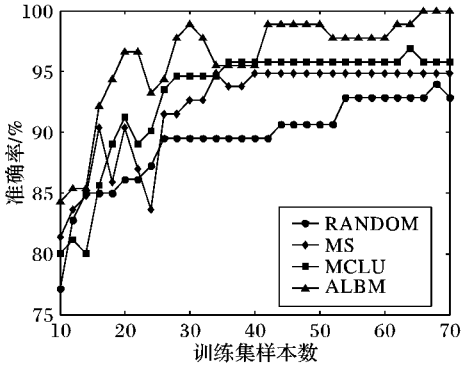


图 1 在 wine 数据集上的分类准确率

为了测试流形主动学习算法在遥感图像分类邻域的应用,本文使用高光谱图像数据集 Indiana's Indian Pine,该图像包含 220 个基带,16 个不同的类别;数据集中含有 10 366 个标记样本。由于需要剔除湖水影响的类别,在实验中,删除 100 个标记的样本点,最终形成以 13 个类别和 10 266 个标记样本点的数据集。在本实验中,选出 7000 个样本作为训练集和候选样本集,3 266 个样本作为测试集。

对高光谱数据集而言,初始训练集含有 130 个标记样本,每次迭代时选取 20 个样本加入训练样本集,直到分类结果稳定时结束迭代。为了减小初始训练集的随机性对分类结果的影响,主动学习过程重复实验 20 次,最终的分类结果取其平均值,如图 2 所示。从中可以看出,本文的 ALBM 算法的分类准确率高于其他的算法;此外,ALBM 算法分类结果收敛时只使用了 1900 个样本,远远少于总样本数量,因此,算法效率高。

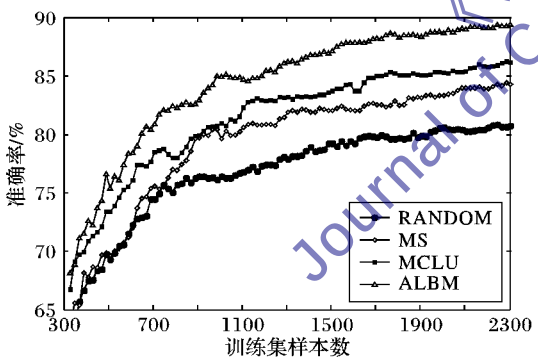


图 2 在高光谱图像数据集上的分类准确率

由表 1 可以看出,在迭代开始阶段,样本数量少,随机抽样方法的 kappa 系数高于边缘抽样。但是经过几次迭代后,MS 和 MCLU 算法要优于随机抽样。因为上述两种算法所选样本的质量高于随机抽样。另外,由于 MS 和 MCLU 算法都是考虑样本到决策边界的距离,所以 kappa 系数很相近。值得注意的是,从迭代过程开始直到结束,本文的 ALBM 算法的 kappa 系数总是优于其他主动学习算法。

## 3 结语

为了有效地解决遥感图像分类问题,本文提出了基于流形主动学习算法,充分考虑到样本数据的空间结构,最大限度地利用标记样本的信息。通过与现有的主动学习算法比较,实验结果表明该算法获得了更高的分类准确率和更好的分类一致性。

鉴于该算法的计算复杂度问题,尚不能适用于大规模数据集,下一步工作将研究如何在预处理阶段选取最具代表性的样本用于训练模型,减小计算复杂度。

表 1 各主动学习算法的 kappa 系数对比

| $n$ | kappa 系数 | | | |
|---|---|---|---|---|
| | RANDOM | MS | MCLU | ALBM |
| 300 | 0.6823 | 0.6616 | 0.6717 | 0.6961 |
| 450 | 0.7337 | 0.7088 | 0.7172 | 0.7477 |
| 600 | 0.7603 | 0.7476 | 0.7500 | 0.7679 |
| 750 | 0.7744 | 0.7807 | 0.7834 | 0.7866 |
| 900 | 0.7802 | 0.7889 | 0.7895 | 0.7926 |
| 1050 | 0.7923 | 0.7968 | 0.7981 | 0.8003 |
| 1200 | 0.7930 | 0.8034 | 0.8051 | 0.8094 |
| 1350 | 0.8022 | 0.8059 | 0.8069 | 0.8199 |
| 1500 | 0.8101 | 0.8105 | 0.8148 | 0.8203 |
| 1650 | 0.8162 | 0.8180 | 0.8190 | 0.8263 |
| 1800 | 0.8189 | 0.8328 | 0.8337 | 0.8390 |

**参考文献:**

[1] HASTIE T, TIBSHIRANI R, FRIEDMAN J. The elements of statistical learning: data mining, inference, and prediction [M]. 2nd ed. New York: Springer, 2009.

[2] BOSER B E, GUYON I M, VAPNIK V N. A training algorithm for optimal margin classifiers [C]// COLT '92: Proceedings of the Fifth Annual Workshop on Computational Learning Theory. New York: ACM, 1992: 144 - 152.

[3] HAYKIN S O. Neural networks and learning machines [M]. 3rd ed. Cambridge: Prentice-Hall, 2008.

[4] SETTLES B. Active learning literature survey, Computer Science Technique Report 1648[R]. Madison, WI: University of Wisconsin-Madison, 2010.

[5] TONG S, KOLLER D. Support vector machine active learning with applications to text classification [J]. Journal of Machine Learning Research, 2002, 2: 45 - 66.

[6] OLSSON F. A literature survey of active machine learning in the context of natural language processing, SICS Technical Report T2009:06 [R]. Kista, Sweden: Swedish Institution Computer Science, 2009.

[7] 蒋华, 戚玉顺. 基于球结构支持向量机的多标签分类的主动学习[J]. 计算机应用, 2012, 32(5): 1359 - 1361.

[8] de SA V R. Learning classification with unlabeled data [C]// Proceedings of Advances in Neural Information Processing Systems. San Francisco: Morgan Kaufmann, 1994: 112 - 119.

[9] (美) CRISTIANINI N, SHAWE-TAYLOR J. 支持向量机导论[M]. 李国正, 王蒙, 曾华军, 译. 北京: 电子工业出版社, 2004.

[10] 张伟, 柳先辉, 丁毅, 等. 基于支持向量回归的多时间序列自回归方法[J]. 计算机应用, 2012, 32(9): 2508 - 2511.

[11] HADSELL R, CHOPRA S, LECUN Y. Dimensionality reduction by learning an invariant mapping [C]// CVPR '06: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Washington, DC: IEEE Computer Society, 2006, 2: 1735 - 1742.

[12] 石陆魁, 张军, 宫晓腾. 基于邻域保持的流形学习算法评价模型[J]. 计算机应用, 2012, 32(9): 2516 - 2519.

[13] 张瑞丽, 张继福. 基于 w-距离均值的模糊聚类算法[J]. 计算机应用, 2012, 32(7): 1978 - 1982.

[14] 邵超, 张慧娟. 应用于不完整流形的 ISOMAP 算法[J]. 计算机应用, 2012, 32(7): 1987 - 1990.

[15] 易淼, 刘小兰. 基于相对变换的半监督分类算法[J]. 计算机应用, 2011, 31(10): 2793 - 2795.