软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2014,25(5):997−1013 [doi: 10.13328/j.cnki.jos.004452]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

双层随机游走半监督聚类*

何 萍, 徐晓华, 陆 林, 陈 崚

(扬州大学 信息工程学院 计算机系,江苏 扬州 225009)

通讯作者: 徐晓华, E-mail: arterx@gmail.com

摘 要: 半监督聚类旨在根据用户给出的必连和不连约束,把所有数据点划分到不同的簇中,从而获得更准确、更加符合用户要求的聚类结果.目前的半监督聚类算法大多数通过修改已有的聚类算法或者结合度规学习,使聚类结果与点对约束尽可能地保持一致,却很少考虑点对约束对周围无约束数据的显式影响程度.提出一种由在顶点上的低层随机游走和在组件上的高层随机游走两部分构成的双层随机游走半监督聚类算法,其中,低层随机游走主要负责计算选出的约束顶点对其他顶点的影响范围和影响程度,称为组件;高层随机游走则进一步将各个点对约束以自适应的强度在组件上进行约束传播,把它们在每个顶点上的影响综合在一个簇指示矩阵中.UCI 数据集和大型真实数据集上的实验结果表明,双层随机游走半监督聚类算法比其他半监督聚类算法更准确,也比较高效.

关键词: 半监督聚类;点对约束;随机游走;组件;影响扩散

中图法分类号: TP181

中文引用格式: 何萍,徐晓华,陆林,陈崚.双层随机游走半监督聚类.软件学报,2014,25(5):997−1013. http://www.jos.org.cn/1000-9825/4452.htm

英文引用格式: He P, Xu XH, Lu L, Chen L. Semi-Supervised clustering via two-level random walk. Ruan Jian Xue Bao/Journal of Software, 2014,25(5):997−1013 (in Chinese). http://www.jos.org.cn/1000-9825/4452.htm

Semi-Supervised Clustering via Two-Level Random Walk

HE Ping, XU Xiao-Hua, LU Lin, CHEN Ling

(Department of Computer Science, College of Information Engineering, Yangzhou University, Yangzhou 225009, China)

Corresponding author: XU Xiao-Hua, E-mail: arterx@gmail.com

Abstract: Semi-Supervised clustering aims to partition the data points into different clusters based on the user-specified must-link and cannot-link constraints. The current semi-supervised clustering algorithms either modify the clustering methods or combine the metric learning approaches to adapt the clustering result as consistent with the pairwise constraints as possible. However, few of them try to explicitly compute the degrees of influence that each pairwise constraint exerts on the unconstrained data points. This paper proposes a semi-supervised clustering algorithm via a two-level random walk, which is composed of a lower-level random walk on vertices and a higher-level random walk on components. The lower-level random walk is responsible for computing the influence range of every vertex constrained by a pairwise constraint. This information is encapsulated in an intermediate structure called “component”. The higher-level random walk further propagates the pairwise constraints on the components with adaptive strength, followed by the integration of all the constraint influence into a cluster indicating matrix. The experiments on UCI database and large real-world data sets demonstrate that, compared with other semi-supervised clustering algorithms, the proposed method not only produces more satisfactory clustering results but also exhibits good efficiency.

Key words: semi-supervised clustering; pairwise constraint; random walk; component; influence expansion

半监督聚类,又称为约束聚类,是指在给定一组点对约束的前提下对已知数据进行聚类.通常,点对约束包

* 基金项目: 国家自然科学基金(61003180, 61070047, 61103018); 江苏省自然科学基金(BK2010318); 江苏省教育厅自然科学基金(13KJB520026, 09KJB20013); 江苏省研究生科技创新计划(CXLX12_0917); 扬州大学新世纪人才计划

收稿时间: 2012-04-25; 修改时间: 2012-10-19; 定稿时间: 2013-07-02

括必连约束(must-link constraint)和不连约束(cannot-link constraint)两种类型[1],其中,每个必连约束指定属于同簇的一对数据点,每个不连约束指定属于不同簇的一对数据点.如果我们将数据看作是图上的顶点,那么给定点对约束的半监督聚类就对应着具有边约束的图上学习.一个好的半监督聚类算法应当将相似的数据划分在同簇,而将不相似的数据划分在异簇;与此同时,尽可能地满足给定的点对约束.与传统的无监督聚类相比,半监督聚类充分利用了人为给定的先验信息作为指导,产生出更为准确、更加符合用户要求的聚类结果,因此,近年来吸引了越来越多研究者的关注[2−4],并被成功地应用到名词词组解析、GPS 车道定位、文本数据分析、调控模块挖掘、DBLP 姓名区分和视频对象识别等多个领域中去[5].

早期的半监督聚类算法大多通过修改一些传统的算法,如 *k*-means[6]、所有点对间的最短路径[1]以及高斯混合模型[7]等来寻找能够满足所有约束条件的一个聚类结果.这些算法为了满足增加的约束条件,将解空间缩小为原始可行解集的一个子集,并且大多倾向于使用贪婪方法在原始可行解集中进行搜索,所以它们给出的聚类结果虽然可以满足约束条件,但却往往并不一定是最好的.

半监督聚类的另一个研究思路是进行度规学习(metric learning),即,通过学习一个合适的距离度规,使得必连的数据相互靠近,而不连的数据相互远离[8,9].Bilenko 等人[10]提出了一种结合了约束 *k*-means 和度规学习的混合算法 MPCKmeans,其实验结果显示,该算法的性能超过了任何一种单一的途径.

然而,所有涉及度规学习的算法都至少存在以下几个共同的难题:

- 一是需要足够的监督信息才能学到相对正确的度规.
- 二是需要对所学度规的适用范围做出先验假设——全局度规虽是最为常用的简单度规,却不适用于强调局部结构的流形数据集;局部度规更为精准,可是除了学习代价高昂之外,还要对“局部”的定义做出额外的假设.
- 三是度规学习只针对约束数据的相似度进行优化,因而算法性能对点对约束的不同位置非常敏感.

随着谱方法在无监督聚类领域的风靡,近年来又涌现出了一批基于谱方法的半监督聚类算法.Kamvar 等人[11]提出了第一种半监督的谱聚类算法(spectral learning).它通过将相似度矩阵中必连约束的相似度修改为 1,不连约束的相似度修改为 0,然后对修改后的相似度矩阵进行谱聚类,获得受约束条件引导的聚类结果.可是,这种 1/0 相似度修改策略并不十分合理,因为属于同一个簇的数据点并不一定完全重合,而属于不同簇的数据点也不一定完全无关.

为了避免这种极端的修改策略,Kulis 等人[12]提出了一种更为柔性的半监督聚类算法(SS-kernel-Kmeans).它采用了与 Spectral Learning 基本相同的流程,只是将其中 1/0 相似度修改策略替换为奖励/惩罚(reward/penalty)策略,即在必连约束的相似度上加上一个奖励值,而在不连约束的相似度上扣去一个惩罚值.其优点是涵盖了 Kamvar 等人的 1/0 修改策略,并且可以通过调节奖励值和惩罚值得到更为温和的相似度修改;其缺点在于没有对不连约束的惩罚值设置限制,当预定的惩罚值大于原相似度值时,可能会产生非正定矩阵,导致算法无法收敛.

以上两种半监督谱聚类算法都仅仅依靠修改约束边的相似度来影响最终的聚类结果,没有充分利用数量有限却内涵丰富的监督信息.

为了克服这一弱点,研究者们尝试了各种不同的方法,将约束边的影响扩大到无约束的边上.例如,Yu 等人[13]提出了约束归一化割的目标函数,De Bie 等人[14]修改了拉普拉斯矩阵的特征值空间,Lu 等人[15]采用高斯过程等.但是,这些方法要么无法处理多聚类问题,要么只能处理必连约束.Li 等人[16]结合谱方法和度规学习两种手段,在数据的谱空间中寻找能够尽可能保持数据与点对约束之间一致性的全局度规.他们的算法(CCSR)既可以处理多聚类问题,又可以接受必连约束和不连约束两种不同类型的监督信息.然而正如前面所述的度规学习的缺点,选择全局度规就意味着每个点对约束的影响会被均匀地扩散到所有无约束的边上,这对于平面上的数据集或许尚且合理,但对于强调局部结构多于全局结构的黎曼流形上的数据却并不合适.实际上,真实世界中的数据集通常位于潜在的黎曼流形上,因而点对约束在附近边上施加的影响也往往大于在远处边上的影响.要实现点对约束的局部影响传播,最直观方法的是使用局部度规.MPCKmeans 算法[10]就假设每个簇都有一个独

立的度规,在 *k*-means 聚类过程中渐进优化各个簇的局部度规.但是除了要有足够的数据来支持准确的度规学习之外,MPCKmeans 还在处理一个簇中包含有多种不同度规的数据集时存在假设与实际不符的问题.

本文提出了一种基于双层随机游走的半监督聚类算法,简称为 SCRAWL(semi-supervised clustering via random walk).它可以将点对约束的影响局部地且光滑地传播到周围无约束的边上.

- 对于任意一个给定的点对约束,SCRAWL 首先确定每个约束顶点的传播范围以及它对传播范围内各个无约束顶点的影响程度,将其定义为介于粗粒度的簇和细粒度的顶点之间的一种中间结构,称为组件.
- 然后,SCRAWL 以组件为单位进行自适应的约束传播,将各个点对约束的影响按照无约束边所连接的顶点与有约束顶点之间的相似度,成比例地传播开去——与约束顶点越相近的顶点所连接而成的无约束边,满足与给定约束边相同关系的概率就越大;与约束顶点离得越远的顶点所连接而成的无约束边,受到的传播影响就越小.
- 最后,SCRAWL 综合所有点对约束在不同组件上施加的影响,把顶点的组件指示矩阵与组件的簇指示矩阵两者相乘,作为每个顶点的簇指示矩阵,从而获得最终的聚类结果.

从对监督信息的利用方式来看,以往的半监督聚类算法都仅限于在边上利用已知的点对约束,而 SCRAWL 则是将约束边的信息先转化为受约束的顶点(边→点),然后把约束顶点的的影响扩散到周围无约束的顶点片段上(点→点),最后再将点对约束施加到受影响的点集的边上(点→边),从而实现边→边的局部影响传播,因此是一种"边→点→边"的监督信息利用方式.

本文第 1 节提出 SCRAWL 算法.第 2 节对 SCRAWL 的性能进行评价,并讨论参数选择.最后,第 3 节进行总结.

1 SCRAWL 算法

定义 1(半监督聚类). 给定一个数据集 $X=\{x_1,x_2,\ldots,x_n\}$和一个点对约束集合 $C=C_{=}\cup C_{\neq}$,其中,$C_{=}$为必连约束集合,它的元素 $c_{=}(x_i,x_j)$表示 x_i 和 x_j 属于同一个簇;$C_{\neq}$为不连约束集合,它的元素 $c_{\neq}(x_i,x_j)$表示 x_i 和 x_j 属于不同的簇.半监督聚类是指将 X 划分到 p 个簇中,同时尽可能地满足 C 中的点对约束.

性质 1(约束的传递性[6]). 如果$\exists c_{=}(v_i,v_k)$且$\exists c_{=}(v_k,v_j)$,则$\exists c_{=}(v_i,v_j)$成立;如果$\exists c_{=}(v_i,v_k)$且$\exists c_{\neq}(v_k,v_j)$,则$\exists c_{\neq}(v_i,v_j)$成立.

定义 2(点对约束传递闭包[6]). 根据性质 1 推导而得的扩增的点对约束集称为 C 的传递闭包.

对于任意一个半监督聚类问题,我们将数据集 X 映射到一个无向加权图 $G=(V,E,w)$上,其中,$V=\{v_1,v_2,\ldots,v_n\}$代表顶点集,顶点 v_i 对应着数据 x_i;E 为边集;w 为定义在边集 E 上的相似度函数,它通常对应一个相似度矩阵 W.根据顶点与点对约束之间的关系,V 可被进一步划分为以下 4 个子集:

定义 3(必连约束顶点集). 必连约束顶点集 $V_{c_{=}}$ 由所有被 $C_{=}$约束的顶点构成:

$$V_{c_{=}} \triangleq \{v_i,v_j \mid \exists c_{=}(v_i,v_j)\in C_{=}\} \tag{1}$$

定义 4(不连约束顶点集). 不连约束顶点集 $V_{c_{\neq}}$ 由所有被 $C_{\neq}$约束的顶点构成:

$$V_{c_{\neq}} \triangleq \{v_i,v_j \mid \exists c_{\neq}(v_i,v_j)\in C_{\neq}\} \tag{2}$$

定义 5(约束顶点集). 约束顶点集 V_c 由所有被 C 约束的顶点构成:

$$V_c \triangleq V_{c_{=}} \cup V_{c_{\neq}} \tag{3}$$

定义 6(无约束顶点集). 无约束顶点集 V_u 为约束顶点集的补集:

$$V_u \triangleq V/V_c \tag{4}$$

在相似度函数 w 的设置上,顶点 v_i 和 v_j 的相似度(简写为 $w(i,j)$)应满足以下准则:

准则 1(相似度准则). 相似度值 $w(i,j)$应满足:$\forall(i,j)$,

(i) 非负有限性:$0\leqslant w(i,j)<\infty$;

(ii) 自相似性:$w(i,j)\leqslant w(i,i)$且 $w(i,j)=w(i,i)\Leftrightarrow i=j$;

(iii) 对称性:$w(i,j)=w(j,i)$.

在实际应用中,人们通常将相似度限定在[0,1]之间,得到更为实用的准则 1′:

准则 1′(相似度准则). 相似度值 $w(i,j)$应满足:$\forall(i,j)$,

(i′) 非负有界性:$0\leqslant w(i,j)\leqslant 1$;

(ii′) 自相似性: $w(i,i)=1$ 且 $w(i,j)=w(i,i)\Leftrightarrow i=j$;

(iii) 对称性:$w(i,j)=w(i,i)$.

在半监督聚类中,按照 C 在 E 上提供的信息,我们还需对顶点间的相似度进行相应的修改,使其与 C 保持一致.一个合理的相似度修改策略应符合以下准则:

准则 2(相似度修改准则). 令~表示相似度的修改算子,它应满足:

(i) 同簇增强: $\tilde{w}(i,j)\geqslant w(i,j)$, if $\exists c_{=}(v_i,v_j)$.

(ii) 异簇减弱: $\tilde{w}(i,j)\leqslant w(i,j)$, if $\exists c_{\neq}(v_i,v_j)$.

(iii) 自相容性: $\tilde{w}(i,j)$ 仍满足准则 1′.

目前,有两种已有的相似度修改策略:

- 一种是由 Kamvar 等人[11]提出的 1/0 策略,对应到准则 2 为:

(i) $\tilde{w}(i,j)=1$, if $\exists c_{=}(v_i,v_j)$.

(ii) $\tilde{w}(i,j)=0$, if $\exists c_{\neq}(v_i,v_j)$.

(iii) 满足自相容性.

- 另一种是由 Kulis 等人[12]提出的+/-策略,对应到准则 2 为:

(i) $\tilde{w}(i,j)=w(i,j)+\delta_{=}$, if $\exists c_{=}(v_i,v_j)$.

(ii) $\tilde{w}(i,j)=w(i,j)-\delta_{\neq}$, if $\exists c_{\neq}(v_i,v_j)$.

(iii) 不满足自相容性.

可以看到:第 2 种+/–策略违背了相似度非负的条件,当设置的不连约束惩罚项大于原始的相似度值时,可能会产生非正定的相似度矩阵,影响算法的收敛.另一方面,虽然 1/0 策略从理论上来说完全符合准则 2,但我们认为这种修改过于极端,不符合实际.因为两个顶点虽然属于同簇,但却未必一定重合;两个顶点尽管属于异簇,却也未必全无相似之处.为了克服以上两种方法的缺陷,本文提出一种 q/q^{-1} 策略,通过设置参数 q 给出更为泛化的相似度修改方法.

定理 1(q/q^{-1} 相似度修改策略). 给定一个点对约束集 C,修改后的相似度矩阵为 $\tilde{W}=(\tilde{w}(i,j))_{n\times n}$,其元素:

$$\tilde{w}(i,j)\triangleq\begin{cases}w(i,j)^{q}, & \text{if } \exists c_{=}(v_i,v_j)\\ w(i,j)^{1/q}, & \text{if } \exists c_{\neq}(v_i,v_j)\\ w(i,j), & \text{otherwise}\end{cases}\tag{5}$$

参数 $q\in(0,1]$控制着相似度修改的力度.可以证明,该策略是满足相似度修改准则的.

证明:对应到准则 2,当 $q\in(0,1)$时,$\forall w(i,j)\in(0,1)$:

(i) 因为 $\tilde{w}(i,j)_q\geqslant\tilde{w}(i,j)$, if $\exists c_{=}(v_i,v_j)$,所以满足同簇增强.

(ii) 因为 $\tilde{w}(i,j)_{1/q}\leqslant\tilde{w}(i,j)$, if $\exists c_{\neq}(v_i,v_j)$,所以满足异簇减弱.

(iii) 由情形(i)、情形(ii)可知 $\tilde{w}(i,j)\in[0,1]$,所以满足自相容性.证毕. □

推论 1. Kamvar 等人的 1/0 策略是 q/q^{-1} 策略的一种极限情况下的特例.

证明:当 $q\to 0$ 时:

(i) $\tilde{w}(i,j)\to 1$, if $\exists c_{=}(v_i,v_j)$.

(ii) $\tilde{w}(i,j)\to 0$, if $\exists c_{\neq}(v_i,v_j)$.

(iii) 满足自相容性.证毕. □

这里,我们可以把 q 看作是对约束边原始相似度的维持程度:q 值越小,对约束边原始相似度的维持度越低,换言之,即修改程度越大;q 值越大,对约束边原始相似度的维持度越高,即修改程度越小.当 $q\to 0$ 时,对 W 进行 1/0 策略的极端相似度修改;当 q=1 时,保持原始的相似度矩阵不变.

然而,要充分利用 C 中的监督信息以最大限度地提高半监督聚类的性能,仅仅依靠修改约束边的相似度是远远不够的.在此基础上,SCRAWL 还需要进一步扩大点对约束的影响至周围无约束的边上.其基本思想是:借鉴标签传播[17]这一半监督分类算法,选出一批赋有约束信息的代表顶点,借助顶点层的随机游走确定每个代表点的影响范围及其对无约束顶点上的影响程度,并将其封装在一个称为"组件"的中间结构中;然后,在组件上进行自适应的点对约束传播;最后,结合组件层的随机游走,将所有点对约束经由不同组件在顶点上施加的影响融合在一个顶点的簇指示矩阵中,获得最终的半监督聚类结果.

1.1 低层随机游走

假设图 G 中的每个顶点都是一条时齐马尔可夫链上的不同状态,其中一小部分顶点代表吸收状态,其余顶点皆代表转移状态.当随机游走开始时,每个顶点处都分布有一个粒子,每个粒子在每一步随机游走中以 p_{ij} 的概率从顶点 v_i 移动到顶点 v_j.如果它到达了任意一个吸收状态,就留在原地不动;否则,就继续移动.整个随机游走过程当且仅当所有粒子都被吸收后停止.

在设置随机游走的吸收边界时,我们令代表吸收状态的顶点数为

$$s=\max(s_l,\min(|V_c|,s_u)) \tag{6}$$

s 的下界为 s_l,上界为 s_u,在$[s_l,s_u]$范围内,s 的个数随着约束顶点的个数增加而线性增长.在代表顶点的选择过程中,我们遵循如下规则:

$$V_{c_=}\succ V_{c_\neq}\succ V_u \tag{7}$$

即在 V_c 和 V_u 两者之间,V_c 优先选于 V_u,因为点对约束只能借助约束顶点的影响范围传播到无约束的顶点上.仅当$|V_c|<s_l$ 时,V_u 才会作为备选以确保有足够的顶点可以作为吸收状态.因此,SCRAWL 算法还可以用来处理传统的无监督聚类问题($|V_c|=0$).如果$|V_c|>s_u$,则在 V_c 内部 $V_{c_=}$ 优先选于 $V_{c_\neq}$,原因是我们意在把代表顶点的影响范围合并到不同的簇中,必连约束能够比不连约束为后者提供更为直接的监督信息.在 $V_{c_=}$,$V_{c_\neq}$ 和 V_u 这 3 个子集内部,代表顶点的选择是随机的且唯一的.

令 V_a 表示代表吸收状态的顶点集,V_r 表示代表转移状态的顶点集,它们满足 $V_a\cup V_r=V$.我们对应 V_a 和 V_r 的划分,将顶点间的转移概率矩阵 $P=\tilde{D}^{-1}\tilde{W}$ (其中, $\tilde{D}=diag(\tilde{W}\mathbf{1}_n)$)重新组织为

$$P=\begin{bmatrix}P_{aa} & P_{ar}\\ P_{ra} & P_{rr}\end{bmatrix} \tag{8}$$

其中,P_{aa} 和 P_{rr} 分别是 V_a 和 V_r 内部的转移概率子矩阵,P_{ar} 和 P_{ra} 则是 V_a 和 V_r 之间的相互转移概率子矩阵.

假设 $F=(f_{ij})_{n\times n}$ 表示各个粒子在到达稳态分布时被 s 个不同吸收状态所吸收的概率矩阵,它由两部分构成:

$$F=[F_a^T\ F_r^T]^T.$$

其中,F_a 表示从 V_a 出发的粒子被吸收的概率子矩阵,F_r 表示从 V_r 出发的粒子被吸收的概率子矩阵.

可以证明[17],F_a 和 F_r 的状态转移方程为

$$\begin{cases}F_a^{t+1}=F_a^0\\ F_r^{t+1}=P_{ra}F_a^0+P_{rr}F_r^t\end{cases} \tag{9}$$

因此,F 的收敛解为

$$F^*=\begin{bmatrix}F_a^*\\ F_r^*\end{bmatrix}=\begin{bmatrix}F_a^0\\ (I-P_{rr})^{-1}P_{ra}F_a^0\end{bmatrix} \tag{10}$$

虽然类似的解已经被用于半监督分类领域[17,18],但本文处理的问题与此完全不同.半监督分类是给定部分顶点的类别标签,预测未知顶点的所属类别;而半监督聚类则是给定部分边的相连或不连关系,预测所有顶点的簇划分.如果我们将半监督分类看作是具有点约束的图上学习问题,那么半监督聚类就是具有边约束的图上学

习问题.相比之下,后者信息量更少,预测难度更大.其次,状态指示矩阵 F 所表达的意义和初始状态的设置方法也不同.在半监督分类中,F 指示各个顶点属于不同类别的隶属度,标签顶点的初始状态满足 $f_{ij}=1$ iff $y_i=c_j$(第 j 类),否则 $f_{ij}=0$;而在 SCRAWL 的低层随机游走中,F 表示各个顶点受到不同代表顶点影响的程度,F_a^0 的设置依赖于对代表顶点状态间相互关系的不同假设.为简单起见,本文假设各个代表顶点的状态相互独立,互不影响,即 $F_a^0=I$.

因此,由公式(11)可知,F 的整体收敛解为

$$F=\begin{bmatrix}F_a\\F_r\end{bmatrix}=\begin{bmatrix}I\\(I-P_{rr})^{-1}P_{ra}\end{bmatrix} \tag{11}$$

它的元素 f_{ij} 表示顶点 v_i 受到第 j 个代表顶点 v^j 的影响程度,$\forall v^j\in V_a$ 的影响范围包括所有 $f_{ij}>0$ 的顶点 v_i.

1.2 高层随机游走

定义 7(组件). 令 T 表示组件的集合,$|T|$表示组件的个数.第 j 个组件 T_j 由代表第 j 个吸收状态的顶点和受它影响的其他顶点片段构成:

$$T_j=\{f_{ij}v_i|f_{ij}>0\} \tag{12}$$

因此,我们称 F 为顶点的组件指示矩阵,F 的元素 f_{ij} 指示了顶点 v_i 对 T_j 的隶属度.显然,组件的个数等同于吸收状态的个数,$|T|=s$.

定理 2. 组件间的两两相似度矩阵为

$$W_c=F^T\tilde{W}F \tag{13}$$

证明:$W_c=(w_c(\alpha,\beta))_{|T|\times|T|}$,其元素:

$$w_c(\alpha,\beta)=\sum_{i=1}^{n}\sum_{j=1}^{n}f_{i\alpha}f_{j\beta}\tilde{w}_{ij}=f_{\cdot\alpha}^T\tilde{W}f_{\cdot\beta} \tag{14}$$

它等于组件间的顶点相似度之和,每对顶点相似度的权重 $f_{\cdot\alpha}f_{\cdot\beta}$ 等于它们属于指定组件对的概率.证毕. □

定义 8(归一化的组件相似度矩阵). 令 W_c 的度数矩阵为 $D_c=diag(W_c\mathbf{1}_{|T|})$,则归一化的组件相似度矩阵为

$$\bar{W}_c=D_c^{-\frac{1}{2}}W_cD_c^{-\frac{1}{2}} \tag{15}$$

性质 2. 归一化的组件相似度矩阵 $\bar{W}_c$ 的元素值 $\bar{w}_c(\alpha,\beta)\in[0,1]$.

在此基础上,我们进一步使用前面提出的 q/q^{-1} 相似度修改策略,将 C 直接施加到 $\bar{W}_c$ 上,借助约束顶点所属的不同组件把每个点对约束的影响自适应地传播到其他无约束的边上.假设修改后的归一化组件相似度矩阵为 $\tilde{W}_c=(\tilde{w}_c(\alpha,\beta))_{|T|\times|T|}$,其元素:

$$\tilde{w}_c(\alpha,\beta)=\begin{cases}\tilde{w}_c(\alpha,\beta)^{q_{\alpha\beta}}, & \text{if } \exists c_=(v^\alpha,v^\beta)\\ \tilde{w}_c(\alpha,\beta)^{1/q_{\alpha\beta}}, & \text{if } \exists c_{\neq}(v^\alpha,v^\beta)\\ \tilde{w}_c(\alpha,\beta), & \text{otherwise}\end{cases} \tag{16}$$

其中,v^α和 v^β分别为组件 T_α和 T_β的代表顶点,相似度的修改强度由 T_α内的平均顶点相似度 q_α和 T_β内的平均顶点相似度 q_β的均值自适应地确定.如果一个组件内的无约束顶点与代表顶点非常相似,那么关于该代表顶点的点对约束很可能也同样适用于组件内的其他无约束顶点;反之,如果组件内的无约束顶点与代表顶点的相似度很低,那么更安全的方法是维持原始组件间的相似度不变,以免把监督信息错误地传播到不相关的顶点上.换言之,置信度越高的组件受到的约束影响越大,置信度越低的组件受到的约束影响越小.

定义 9(邻接矩阵). 令 $\tilde{W}^I=(\tilde{w}_{ij}^I)_{n\times n}$ 为 $\tilde{W}$ 对应的邻接矩阵,它将 $\tilde{W}$ 的所有非 0 元素都替代为 1:

$$\tilde{w}_{ij}^I=\begin{cases}1, & \text{if } \tilde{w}_{ij}>0\\ 0, & \text{otherwise}\end{cases} \tag{17}$$

定理 3. 组件间边的置信度矩阵为

$$W_e=F^T\tilde{W}^IF \tag{18}$$

证明:$W_e=(w_e(\alpha,\beta))_{|T|\times|T|}$,其元素:

$$w_e(\alpha,\beta)=\sum_{i=1}^{n}\sum_{j=1}^{n}f_{i\alpha}f_{j\beta}\tilde{w}_{ij}^{I}=\sum_{\substack{i,j=1\\ \tilde{w}_{ij}>0}}^{n}f_{i\alpha}f_{j\beta} \tag{19}$$

即所有顶点对属于指定组件对的概率之和.证毕. □

定理 4. 所有组件内部的平均顶点相似度所购成的向量为

$$\zeta=diag(W_c)\ ./\ diag(W_e) \tag{20}$$

其中,./表示对应的元素相除.

证明:$\zeta=(\zeta_\alpha)_{|T|\times1}$,其元素:

$$\zeta_\alpha=\frac{\sum\limits_{\substack{i,j=1\\ \tilde{w}_{ij}>0}}^{n}f_{i\alpha}f_{j\alpha}\tilde{w}_{ij}}{\sum\limits_{\substack{i,j=1\\ \tilde{w}_{ij}>0}}^{n}f_{i\alpha}f_{j\alpha}}=\frac{\sum\limits_{i=1}^{n}\sum\limits_{j=1}^{n}f_{i\alpha}f_{j\alpha}\tilde{w}_{ij}}{\sum\limits_{i=1}^{n}\sum\limits_{j=1}^{n}f_{i\alpha}f_{j\alpha}\tilde{w}_{ij}^{I}}=\frac{f_{\cdot\alpha}^{T}\tilde{W}f_{\cdot\alpha}}{f_{\cdot\alpha}^{T}\tilde{W}^{I}f_{\cdot\alpha}}=\frac{W_c(\alpha,\alpha)}{W_e(\alpha,\alpha)} \tag{21}$$

定理 4 证毕. □

性质 3. 由于 $\tilde{W}$ 中每对顶点相似度的值都在[0,1]范围之内,因此平均顶点相似度ζ的值也在[0,1]范围之内.

根据组件内的平均顶点相似度ζ,我们使用一个 Logistic 函数自适应地确定各个点对约束在不同组件间边上的传播强度:

$$q_{\alpha\beta}=q_0+(1-q_0)\frac{1}{1+e^{\gamma(\zeta_\alpha+\zeta_\beta-1)/2}} \tag{22}$$

其中,$q_0\in[0,1]$作为 $q_{\alpha\beta}\in[q_0,1]$的下界以避免极端的相似度修改,$\gamma>0$ 则控制曲线从最大值到 q_0 的弯曲程度.当$\zeta_\alpha,\zeta_\beta\to0$ 时,T_α和 T_β的顶点平均相似度都很低,不足以用单个约束顶点 v^α和 v^β来代表它们的组件.此时,$q_{\alpha\beta}\to1$,保留组件间的原始相似度.当$\zeta_\alpha,\zeta_\beta\to1$ 时,T_α和 T_β的顶点平均相似度都很高,说明 T_α和 T_β中的顶点很有可能与 v^α和 v^β属于相同的簇,关于 v^α和 v^β的点对约束也很可能适用于组件中的其他顶点.此时,$q_{\alpha\beta}\to q_0$,必连约束组件间的相似度被放大 q_0 幂次,而不连约束组件间的相似度则被缩小为 q_0^{-1} 幂次.当$\zeta_\alpha\to0$ 而$\zeta_\beta\to1$(或反之)时,其中一个组件的置信度很高,另外一个组件的置信度很低,此时,我们取其均值$(\zeta_\alpha+\zeta_\beta)/2$,以大约$(1+q_0)/2$ 的中等强度进行约束传播.

经过点对约束在组件上的自适应传播后,组件间的转移概率矩阵为 $P_c=\tilde{D}_c^{-1}\tilde{W}_c$,其中, $\tilde{D}_c=diag(\tilde{W}_c\mathbf{1}_{|T|})$.

Melia 等人[19]证明了最小化不同组件簇之间的相互转移概率等价于最小化它们的归一化割.已知归一化割的一个近似最优解是由 P_c 的 p 个最大特征值对应的特征向量所构成的[19]:

$$U=[u_1u_2\ldots u_p] \tag{23}$$

其中,$u_1,\ldots,u_p$ 满足 $P_cu_i=\lambda_iu_i$,且$\lambda_1\geqslant\ldots\geqslant\lambda_p$.因为 U 的第 i 行指示了第 i 个组件的所属簇,我们称 U 为组件的簇指示矩阵.

通过将顶点的组件指示矩阵 $F_{n\times|T|}$和组件的簇指示矩阵 $U_{|T|\times p}$ 相乘,我们可获得关于顶点的簇指示矩阵:

$$G_{n\times p}=FU \tag{24}$$

并且 G 的第 i 行指示第 i 个顶点 v_i 的所属簇.如果我们把 U 看作是点对约束施加到组件上的结果,那么 G 就是把点对约束的影响进一步按照顶点的隶属度光滑地传播到每条无约束边上,并且综合各种影响的结果.为了确定每个顶点被划分到哪个簇中,我们采用一种常用的行归一化后处理技术[20],即将 G 的行向量投影到一个单位超球面上:

$$\tilde{G}=D_G^{-1/2}G \tag{25}$$

其中,$D_G=diag(diag(GG^T))$.然后,用 k-means 将 $\tilde{G}$ 的 n 个行向量划分到 p 个簇中.

1.3 时间复杂度

在 SCRAWL 中,推导点对约束传递闭包以及修改顶点相似度矩阵的预处理时间复杂度为 $O(|C|)$.低层随机游走的时间复杂度由组件指示矩阵 F 的计算占主导.虽然我们已经给出了 F 的闭合解(公式(11)),但对于大数据集而言,更为省时的方法是进行迭代计算(公式(9)).假设在给定的稀疏相似度矩阵中,每个顶点只与其最近的 k 个近邻顶点相连接,则 P_{rr} 的每一行最多只含有 k 个非零元素.令 $t_{\max}$ 表示迭代的最大次数,迭代计算 F_r 的时间复杂度为 $O((n-|T|)k|T|t_{\max})$.考虑到 $t_{\max}$ 和 k 都是用户指定的常数,它们可以直接被移除而不影响时间复杂度 $O((n-|T|)|T|)$.另一方面,在 SCRAWL 的高层随机游走中,组件相似度的计算耗费 $O(|T|nk)$的时间复杂度,而 P_c 的特征值分解需要 $O(|T|^3)$的时间复杂度.在处理大型真实数据集时,我们发现$|T|$和$|C|$与 n 相比是如此之小以至于完全可以被忽略,因此使得算法的时间复杂度被降为近线性 $O(n)$.如果再加上稀疏相似度矩阵的构建复杂度 $O(n^2)$,则 SCRAWL 的完整时间复杂度为 $O(n^2)$.

2 实 验

在本节中,我们将 SCRAWL 算法与 Spectral Learning(SL)[11],SS-Kernel-Kmeans(SSKK)[12],Constrained Clustering with Spectral Regularization(CCSR)[21]以及 Metric Pairwise Constrained Kmeans(MPCK)[10]这 4 种最为相关的半监督聚类算法在 UCI 数据集和大型真实数据集上进行了全面的评价和比较.首先,我们将介绍用于算法测试的 12 个数据集;接着,我们给出包括算法实现、参数设置以及评价标准在内的实验设计方案;然后,我们对 5 种半监督聚类算法在各个数据集上的学习曲线进行评价,并根据各个参数对算法性能的影响提供最优参数的选择方案;最后,我们将分析算法的时间复杂度,并比较它们的可扩放性能.

2.1 实验数据

表 1 列举了 12 个测试数据集的基本信息.它包含了 6 个 UCI 数据集和来源于 4 个不同应用领域的 6 个大型真实世界数据集.它们中的大多数都被已有的半监督聚类算法测试过,其中,tissue,parkinsons,statlog 和 breast 都是医学数据集,iris 是植物数据集,ionosphere 是物理数据集,20Newsgroups 和 TDT2 corpus 是文本文档数据集,MNIST 是手写数字 0~9 的光学字符识别数据集,Letter 和 ISOLET 是 26 个英文字母的数据集以及 CMU PIE 为人脸数据库.

Table 1 Summary of the test data sets

表 1 测试数据集汇总

UCI 数据集				真实世界数据集			
数据集	样本数	维数	类别数	数据集	样本数	维数	类别数
tissue	106	9	6	20Newsgroups	2 774	24 253	3
iris	150	4	3	TDT2	6 146	36 771	5
parkinsons	195	22	2	MNIST	6 000	784	10
statlog	270	13	2	Letter (A~F)	4 639	16	6
ionosphere	351	34	2	ISOLET	7 797	617	26
breast	683	9	2	PIE	2 856	1 024	68

在预处理中,我们将 tissue,parkinsons 和 statlog 的属性值缩放到[−1,1]区间内,移除了 breast 的缺失数据和第 1 个属性值(ID 号).我们采用 20Newsgroups 处理最少的 20news-bydate 数据集,并从中选出 3 种完全不同的类别(alt.atheism,rec.sport.baseball,sci.space);使用 TDT2 原数据集中最大的 5 个类别,移除同时出现在两个或多个类别中的文档;将 20Newsgroups 和 TDT2 数据集的属性值归一化为 TFIDF 表示.此外,我们从原始 MNIST 训练数据集的每个类中随机选出 600 个数据,构成一个大小为 6 000 的 MNIST 子数据集,并将其属性值缩放到[−1,1]范围内;选取原 Letter 数据集中的前 6 个字母(A~F);并使用 CMU PIE 中一个姿势为 C27(近正面)的子数据集[22],用于算法的性能测试.

2.2 实验设计

我们使用 Matlab 实现了包括 SCRAWL,SL,SSKK 和 CCSR 在内的 4 种基于图的半监督聚类算法,同时采用由 Basu 等人提供的 MPCK 算法的 Java 实现.对于小型的 UCI 数据集和大型的真实世界数据集,我们根据数据的真实类别,产生至少 10 组不同数目的点对约束.对于每组不同数目的点对约束,我们随机产生 50 种不同的实现,用于算法的平均性能评价.数据集的大小和聚类任务的困难程度共同决定了在每个数据集上提供的点对约束个数.

我们为 SCRAWL,SL,SSKK 和 CCSR 这 4 种基于图的半监督聚类算法在 UCI 数据集上构建全连通图,而在真实世界数据集上构建 20 最近邻(20NN)稀疏图.除了在 20Newsgroups 和 TDT2 数据集上用余弦函数计算文档的相似度以外,其余图的相似度都是基于高斯径向基函数计算的:

$$w_{ij}=\mathrm{e}^{-\frac{\|v_i-v_j\|^2}{2\sigma^2}} \tag{26}$$

对于每组给定的点对约束,我们在$\{2^{-25/5},2^{-24/5},\ldots,2^{24/5},2^{25/5}\}$区间内寻找最优的图构建尺度参数$\sigma$.在算法参数设置方面,SSKK 的奖励和惩罚权重设置为 $n/(p|C|)$,CCSR 的谱空间嵌入维数默认为 D=15.SCRAWL 的顶点相似度矩阵修改参数为 q=0.02,吸收状态个数的上下限为 $s_u=\lceil 0.1n\rceil$和 $s_l=p$;自适应约束传播强度 $q_{\alpha\beta}$的下限参数和弯曲度为 $q_0=q,\gamma=1/q$;在组件指示子矩阵 F_r 的迭代计算中,初始状态 $F_r^0=0$,迭代的最大次数 $t_{\max}=300$,收敛的阈值被设为 $\varepsilon\|F_r^{t-1}\|_F$,其中,$\|F_r^{t-1}\|_F$ 则是每次迭代时 F_r 的上一时间步状态**,$\varepsilon=2^{-52}$是从 1 到下一个 double 精度值的距离.

为了对 5 种半监督聚类算法进行公正的评价,我们使用 Modified Rand Index 评价标准[12]:

$$Rand\ Index=\frac{TP+TN-|C|}{0.5n(n-1)-|C|} \tag{27}$$

TP(true positive)是指同类的点对被划分到相同簇的数目,TN(true negative)是指不同类的点对被划分到不同簇的数目,$|C|$是指点对约束的个数.分子和分母都减去$|C|$的原因是为了排除点对约束的影响,更有利于半监督聚类的评价.以下实验都在一台 Mac 2GHz Intel Core 2 Duo 2GB RAM 的 MATLAB R2010a 环境下执行.

2.3 实验结果

2.3.1 UCI 数据集

图 1 演示了 5 种半监督聚类算法在 6 个 UCI 数据集上的 Rand Index 学习曲线.

从图中我们可以看到:SCRAWL 表现出最佳的聚类性能.虽然一开始 SCRAWL 的学习曲线在数目较小的点对约束上落后于 CCSR,但是随着约束个数的增加,它迅速超过了其他几种算法并逐渐拉开差距;相反,尽管 CCSR 在初始的点对约束上表现出优越的聚类性能,但是当约束条件的数目成倍增长后,CCSR 的性能却很少有提高.这是因为随着约束个数的增加,用简单的线性变换使有限维(D=15)谱空间嵌入满足这些约束条件的难度也越来越大,故聚类结果很难改善.相比之下,SSKK 虽然在图中有些数据集上表现良好,但却在另外一些数据集上表现令人失望.究其原因,是因为奖励和惩罚权重的设置不适合所有的数据集.此外,SL 学习曲线的上升趋势与 SCRAWL 颇为相似,但是由于缺少约束传播的机制,它的聚类性能以较大的差距落后于 SCRAWL.另外,MPCK 不同于其他 4 种基于图的半监督聚类算法,它的学习曲线首先出现一个下降的低谷,然后再慢慢上升.这主要是因为一开始点对约束较少,从中学习到的参数并不十分可靠,但后来,越来越多的约束条件逐渐改进了这一状况[10],因而算法的性能也随之上升.

** 我们在实验中发现:大多数情况下,F_r 收敛实际所需的迭代次数远小于我们设置的最大迭代次数.即便在少数情况下,该稳态分布在达到最大迭代次数时仍没有收敛,对算法的性能也不会有很大的影响[23].

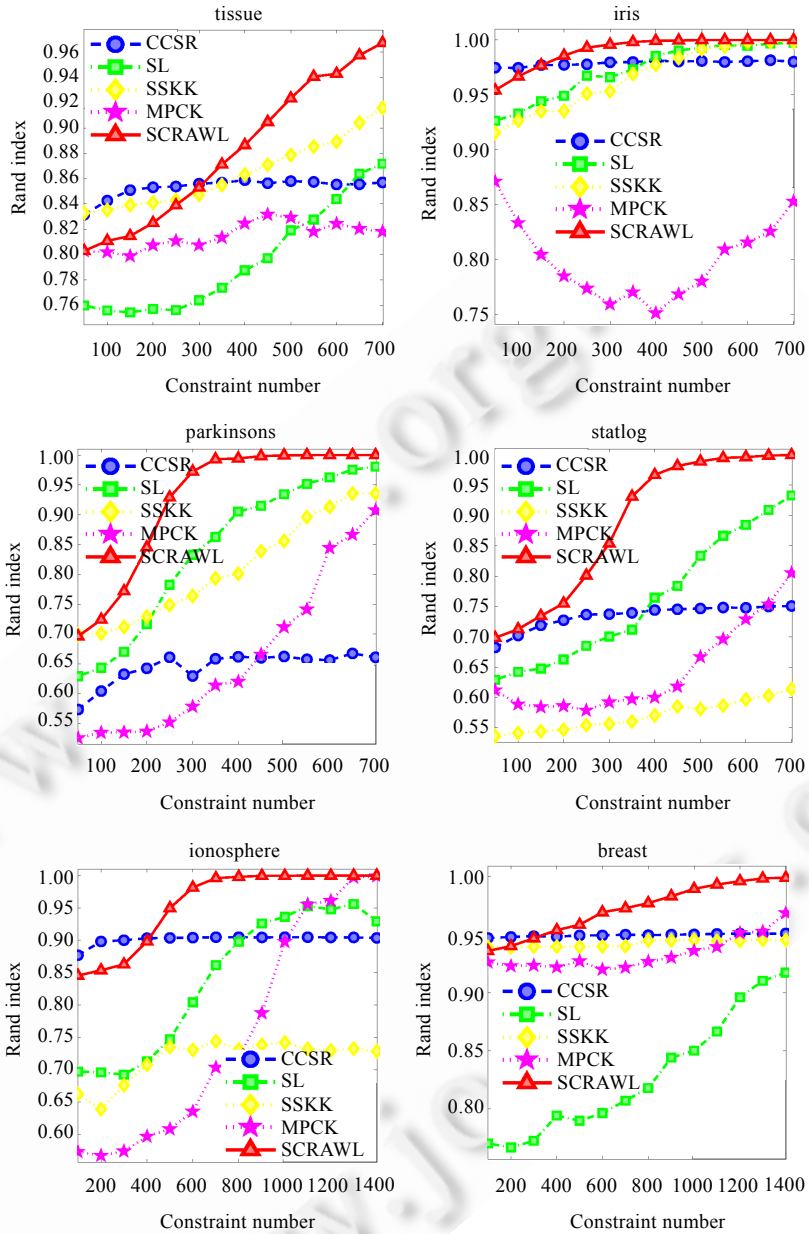


Fig.1 Learning curves of the five semi-supervised clustering algorithms on UCI data sets

图 1 UCI 数据集上 5 种半监督聚类算法的学习曲线

2.3.2 真实世界数据集

图 2 演示了 5 种半监督聚类算法在 6 个真实世界数据集上的 Rand Index 学习曲线.MPCK 没有出现在 TDT2 的比较中,原因是该数据集的维数很高(D=36771),MPCK 无法在合理的时间内学习到完整的度规.

从图中我们可以看出,SCRAWL 仍然表现出 5 种算法中最佳的聚类性能.相比之下,CCSR 性能愈加下降,这或许与图构建方式的变化有关:在 UCI 数据集上,我们为各种算法构建的是全连通图,而在大型真实数据集上,构建的却是 20 最近邻稀疏图,显然,后者对约束传播的局部性具有更严格的要求.由于 CCSR 全局而且统一地改变整个特征空间来满足每一个点对约束,SCRAWL 却将每个点对约束的影响局部且成比例地维持在相应的组

件范围之内,因此,后者能够产生更好的聚类结果.另外,MPCK 的学习曲线极为平坦,其主要原因在于提供的点对约束的增长对于大型数据集而言影响甚微,致使其学到的度规没有实质性变化.

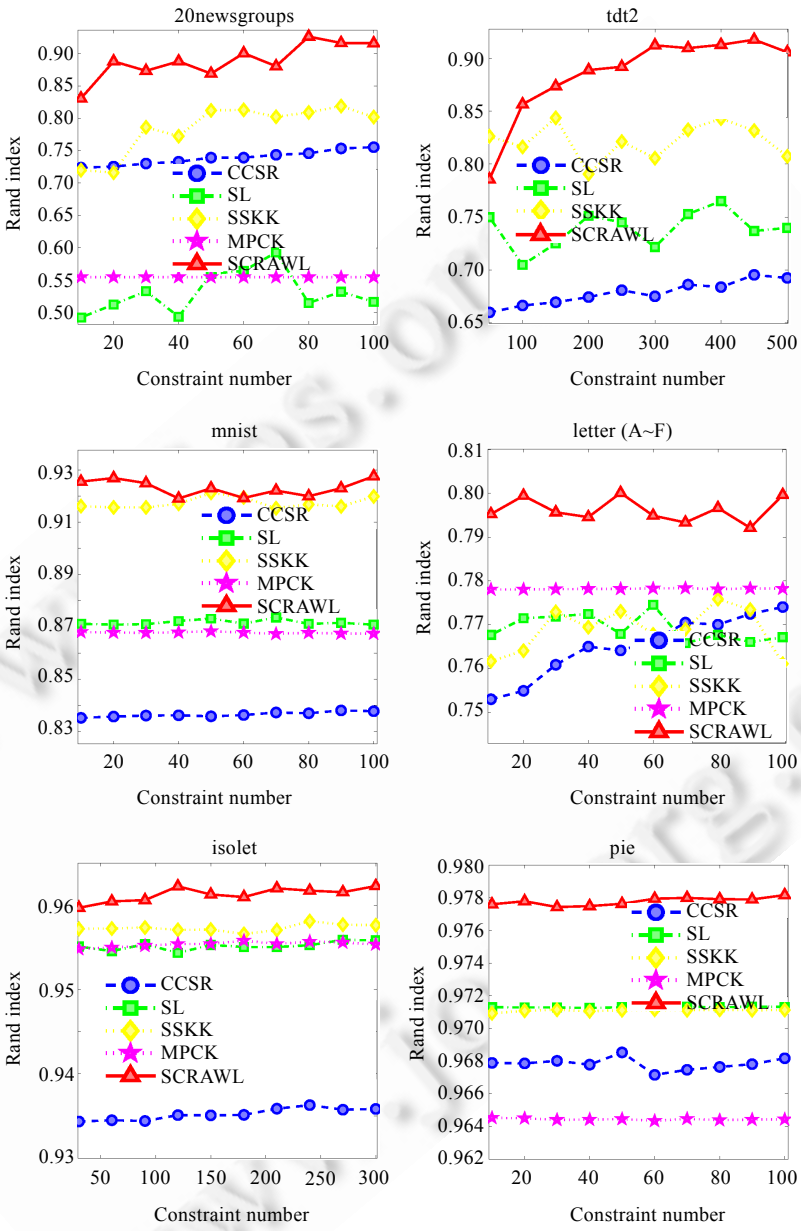


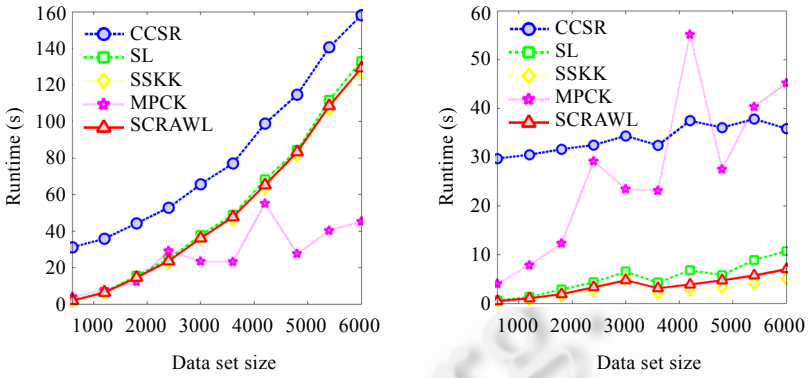
Fig.2 Learning curves of the five semi-supervised clustering algorithms on real-world data sets

图 2 真实世界数据集上 5 种半监督聚类算法的学习曲线

2.3.3 可扩放性

我们从 MNIST 数据集的每个类中以 60 为步长随机选取 60~600 个数据,组成以 600 为步长大小从 600 增长到 6 000 的 10 个独立的 MNIST 子数据集.对每个 MNIST 子数据集,我们基于其真实类别,随机产生 10 个约束条件的 50 种不同实现,然后对 5 种半监督聚类算法进行可扩放性能的评价.

图 3(a)给出了 5 种算法在 MNIST 子数据集组上的完整运行时间比较.



(a) 包含相似度矩阵的计算 (b) 不包含相似度矩阵的计算

Fig.3 Scalability comparison of the five semi-supervised clustering algorithms

图 3 5 种半监督聚类算法的可扩放性比较

由于 CCSR,SL,SSKK 和 SCRAWL 都是基于图的半监督学习算法,因此都要为相似度矩阵的计算耗去 $O(n^2)$ 的时间复杂度,在图上也都呈现出二次的可扩放性曲线.相比之下,MPCK 不依赖于点与点之间的相似度计算,再加上又是用 Java 语言实现(其他算法用 Matlab 实现),故运行时间最短.

其次,在这 4 种基于图的半监督聚类算法中,CCSR 需要花额外的时间在度规学习上,所以其可扩放性曲线比其他 3 种算法曲线的位置要高一些.考虑到相似度矩阵的计算占据了 SCRAWL 等基于图的半监督聚类算法的一大部分运行时间,我们又在图 3(b)中给出了不包含相似度矩阵计算的运行时间比较.从图中我们可以看到:MNIST 子数据集增大 10 倍,SCRAWL 的运行时间也随之增加了大约 10 倍,表现出线性的可扩放性,验证了我们前面的时间复杂度分析.

另外,SL 和 SSKK 一样也表现出线性的可扩放性,CCSR 则在它们的基础上再上移一段距离.这是因为 SL,SSKK 和 CCSR 在对稀疏相似度矩阵进行特征值分解时,使用的是基于 Lanczos 方法的 eigs 函数.虽然 eigs 函数的时间复杂度上限为 $O(n^2)$,但却还依赖于矩阵的具体稀疏度及其最大与第二大特征值之间的差距.

2.4 参数讨论

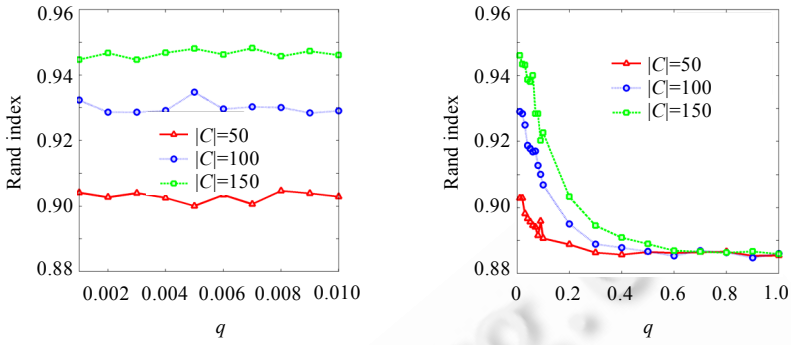
在本文中,SCRAWL 算法共涉及 5 个参数的选择,分别是顶点相似度矩阵 W 的修改强度 q、组件相似度矩阵 W_c 的修改强度下限 q_0 和变化幅度 γ、组件个数的上下限 s_u 和 s_l.

下面,我们以 iris 数据集为例,演示用由简到繁、功能递增的方法逐个确定 SCRAWL 的最优参数,并讨论它们对算法性能的影响.

2.4.1 W 的修改强度 q

为了区分顶点相似度矩阵修改(公式(5))和组件相似度矩阵修改(公式(16))的不同影响,我们固定 q_0=1,保持组件相似度矩阵不变(此时,参数 γ 不起作用),同时,令组件个数等于约束顶点的个数(s=|V_c|),然后对|C|=50,|C|=100 和|C|=150 这 3 种约束条件个数的 50 次不同实现进行性能平均,得到如图 4 所示的 SCRAWL 关于 q 的学习曲线.

从图 4(b)中我们可以看到,当 $q\in$[0.01,1]时,q 值越小,对约束边相似度的修改程度就越大,SCRAWL 的聚类性能也就越佳.而在图 4(a)中,当 $q\in$[0.001,0.01]时,SCRAWL 的聚类性能基本保持不变.这说明 SCRAWL 的顶点相似度矩阵的修改强度 q 在小于某个阈值(如此处为 0.01)时,对算法性能的影响基本不变.



(a) q∈[0.001:0.001:0.01] (b) q∈[0.01:0.01:0.1,0.1:0.1:1]

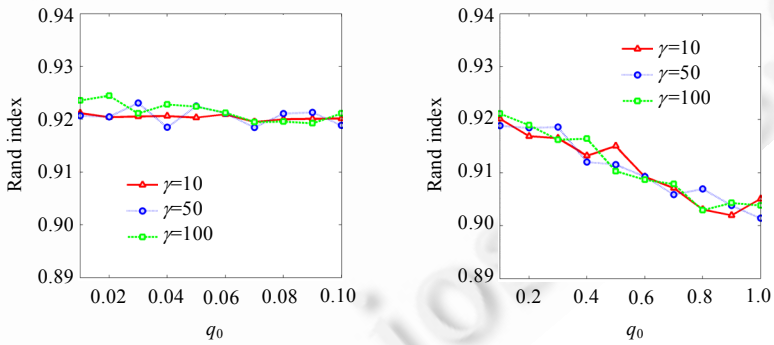
Fig.4 Learning curves of SCRAWL about q on iris data set

图 4 SCRAWL 在 iris 数据集上关于 q 的学习曲线

2.4.2 W_c 的修改强度下限 q_0

根据第 2.4.1 节的分析,我们令 W 的修改强度 q=0.01,s=|V_c|不变,考虑γ={10,50,100}这 3 种不同的情况,分析 W_c 的修改强度下限 q_0 对算法性能的影响.图 5 描绘了|C|=50 时,SCRAWL 关于 q_0 的学习曲线.

可以看出:在图 5(b)中,随着 q_0 从 1 减小到 0.1,点对约束在组件上的传播强度越来越大,SCRAWL 的半监督聚类性能也逐渐提升;而在图 5(a)中,当 q_0∈[0.01,0.1]时,SCRAWL 的半监督聚类性能仅有非常微弱的上升,与 q_0∈[0.1,1]比,几乎可以忽略不计.这说明 SCRAWL 的组件相似度矩阵的修改强度下限 q_0 在小于某个阈值(如,此处为 0.1)时,对算法的半监督聚类性能影响基本不变.另一方面我们也发现:在图 5 中,γ=10,γ=50 和γ=100 这 3 条曲线相互交叠在一起,暗示着参数γ对算法半监督聚类性能的影响不大.



(a) q_0∈[0.01:0.01:0.1] (b) q_0∈[0.1:0.1:1]

Fig.5 Learning curves of SCRAWL about q_0 on iris data set (|C|=50)

图 5 SCRAWL 在 iris 数据集上关于 q_0 的学习曲线(|C|=50)

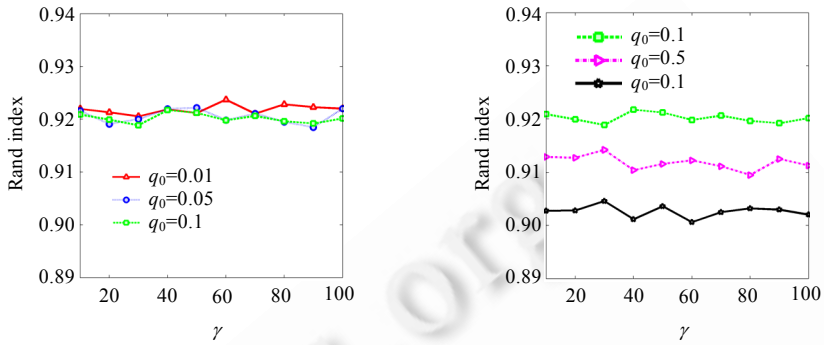
2.4.3 W_c 的修改强度变化幅度γ

基于第 2.4.1 节的分析,我们令 W 的修改强度 q=0.01,维持 s=|V_c|不变,然后考虑 q_0={0.01,0.05,0.1,0.5,1}这 5 种不同情况,考察 W_c 上修改强度的变化幅度参数γ对算法半监督聚类性能的影响.

图 6 描绘了当|C|=50 时,SCRAWL 关于参数γ的学习曲线.

从图 6 中我们可以看到:SCRAWL 的半监督聚类性能对参数γ不敏感.无论 q_0 取什么值,它所对应的γ学习曲线都在均值附近上下波动,没有明显的上升或者下降趋势.在 q_0≤0.1 时(如图 6(a)所示),3 条不同的γ学习曲线在仔细区分下存在很微弱的差距(q_0=0.01≻q_0=0.05≻q_0=0.1),但非常接近,几乎难以区分;而当 0.1≤q_0≤1 时(如图 6(b)所示),3 条γ学习曲线就拉开了明显的距离(q_0=0.1≻q_0=0.5≻q_0=1).其中,q_0=1 的γ学习曲线是性能最差的一条,

因为它没有任何组件层上的点对约束传播(此时γ不起作用).这说明组件上的约束传播是有效的,而且也从另一个侧面验证了第 2.4.2 节的分析结果.即,当 $q_0\in[0.1,1]$时,q_0 值越小,SCRAWL 的半监督聚类性能就越好;而当 $q_0<0.1$ 时,q_0 值对 SCRAWL 的性能影响变化不大.



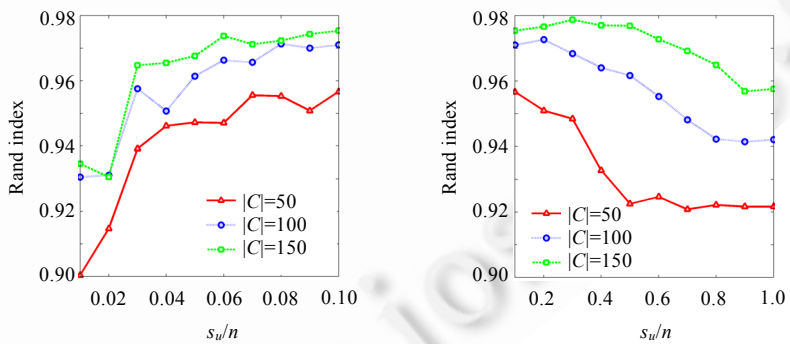
(a) $q_0\in\{0.01,0.05,0.1\}$ (b) $q_0\in\{0.1,0.5,1\}$

Fig.6 Learning curves of SCRAWL about γ on iris data set ($|C|$=50)

图 6 SCRAWL 在 iris 数据集上关于γ的学习曲线($|C|$=50)

2.4.4 组件个数的上限 s_u

结合第 2.4.2 节和第 2.4.3 节的分析,我们令 $q_0=q=0.01$,$\gamma=1/q=100$.考虑到组件的个数不能少于簇的个数,我们将组件个数的下限暂定为 s_l=p,然后讨论组件个数的上限 s_u 对 SCRAWL 半监督聚类性能的影响.图 7 给出了在$|C|$=50,$|C|$=100 和$|C|$=150 这 3 种约束条件个数情况下,SCRAWL 关于 s_u/n 的学习曲线,它的横轴的倒数 n/s_u 指示了每个组件的平均传播范围(单位:顶点数).



(a) $s_u/n\in[0.01:0.01:0.1]$ (b) $s_u/n\in[0.1:0.1:1]$

Fig.7 Learning curves of SCRAWL about s_u/n on iris data set

图 7 SCRAWL 在 iris 数据集上关于 s_u/n 的学习曲线

在图 7(a)中我们看到:当 s_u/n 从 0.01 增加到 0.1 时,随着组件个数的逐渐增加,每个组件的传播范围不断精化,使得 SCRAWL 的性能逐渐上升.但在图 7(b)中,s_u/n 从 0.1 继续增加到 1 时,SCRAWL 的性能却出现了下降的趋势,原因是,随着组件的个数越增越多,每个组件的平均传播范围变得越来越小,点对约束无法被有效地传播到周围的无约束顶点上,因此 SCRAWL 的性能受到了限制.

2.4.5 组件个数的下限 s_l

我们沿用了第 2.4.4 节中的 $q_0=q=0.01$,$\gamma=1/q=100$ 设置,同时,为了避免组件个数上限的影响,暂定 $s_u=n$(即组件的个数不超过顶点的个数),然后分析组件个数的下限 s_l 对 SCRAWL 聚类性能的影响.图 8 演示了 SCRAWL

在|C|=0(无监督聚类),|C|=1 两种|V_c|<p 的情况下关于 s_l/p 的学习曲线,并将其与基准的无监督谱聚类算法(Ncut)进行比较.图中横轴 s_l/p 指示了每个簇中平均包含的组件个数.

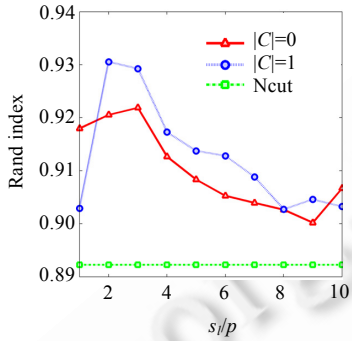


Fig.8 Learning curves of SCRAWL about s_l/p on iris data set

图 8 SCRAWL 在 iris 上关于 s_l/p 的学习曲线

从图 8 中我们看到:s_l/p 的学习曲线呈现出先升后降的趋势,这说明当约束顶点较少时,加入适量的无约束顶点作为组件构建的吸收状态可以较大地提升 SCRAWL 的聚类性能,因为虽然没有约束条件施加在无约束顶点的组件上进行约束传播,但组件本身已经起到了局部聚类、挖掘簇内部潜在结构的效果.另外,从 SCRAWL 的 |C|=0 学习曲线和 Ncut 的基准曲线之间的差距来看,我们完全有理由认为,SCRAWL 不仅是一种优秀的半监督聚类算法,而且也是一种优秀的无监督聚类算法.

至此,我们已经能够通过前面描述的参数选择方法确定一组最优的参数集.我们知道:一组最优的参数集依赖于很多的外部因素,如处理的数据集、提供的点对约束、甚至于选择的评价标准,但是从第 2.3 节中的实验结果来看,虽然我们对所有的数据集采用了一种统一的参数设置,但仍然在整体上取得了较好的聚类结果.再进一步结合第 2.4 节的参数讨论分析来看,SCRAWL 总体上来说并不是一种对参数很敏感的算法,只要用户选择的参数在合适的区间范围内,就能产生较好的半监督或无监督聚类结果.

3 结 论

基于点对约束的半监督聚类,对应到图上是一种具有边约束的学习问题.本文提出了一种双层随机游走的半监督聚类算法 SCRAWL.该方法首先确定每个约束顶点的影响范围;然后,根据每条无约束边所连接的顶点与有约束顶点之间的相似度,将约束边的影响成比例地传播开去;最后,融合所有点对约束的影响,通过计算顶点的簇指示矩阵获得最终的聚类结果.与已有的半监督聚类算法相比,SCRAWL 算法有以下几个方面的优势:

(1) 它将每个点对约束的影响局部地且成比例地扩散到无约束的边上,却无需估计任何度规参数.

(2) 它通过构建组件挖掘出簇内部的潜在结构,有利于用户更好地了解所处理的数据.

(3) 给定预先计算的稀疏相似度矩阵,它可以在线性时间复杂度内处理大型真实数据集;

(4) 它既可以处理二类聚类问题,又可以处理多类聚类问题;既可以利用必连约束,又可以利用不连约束.

大量基于 UCI 数据集和文本文档、手写数字、英文字符以及人脸识别的大型真实数据集实验结果表明:SCRAWL 的聚类性能优于其他几种著名的半监督聚类算法,并且也比较高效.

References:

[1] Klein D, Kamvar SD, Manning CD. From instance-level constraints to space-level constraints: Making the most of prior knowledge in data clustering. In: Proc. of the 19th Int'l Conf. on Machine Learning. Madison: Omnipress, 2002. 307−314.

[2] Xiao Y, Yu J. Semi-Supervised clustering based on affinity propagation algorithm. Ruan Jian Xue Bao/Journal of Software, 2008, 19(11):2803−2813 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/19/2803.htm [doi: 10.3724/SP.J.1001.2008.02803]

[3] Yin XS, Hu EL, Chen SC. Discriminative semi-supervised clustering analysis with pairwise constraints. Ruan Jian Xue Bao/Journal of Software, 2008,19(11):2791−2802 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/19/2791.htm [doi: 10.3724/SP.J.1001.2008.02791]

[4] Wang HJ, Li ZS, Qi JH, Cheng Y, Zhou P, Zhou W. Semi-Supervised cluster ensemble model based on Bayesian network. Ruan Jian Xue Bao/Journal of Software, 2010,21(11):2814−2825 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/3683.htm [doi: 10.3724/SP.J.1001.2010.03683]

[5] Basu S, Davidson I, Wagstaff KL. Constrained Clustering: Advances in Algorithms, Theory, and Applications. Boca Raton: Chapman and Hall/CRC Press, 2008.

[6] Wagsta K, Cardie C, Rogers S, Schroedl S. Constrained *k*-means clustering with background knowledge. In: Proc. of the 18th Int'l Conf. on Machine Learning. Madison: Omnipress, 2001. 577−584.

[7] Shental N, Bar-hillel A, Hertz T, Weinshall D. Computing Gaussian mixture models with EM using equivalence constraints. In: Advances in Neural Information Processing Systems 16. Cambridge: MIT Press, 2003. 465−472.

[8] Xing EP, Ng AY, Jordan MI, Russell S. Distance metric learning, with application to clustering with side-information. In: Advances in Neural Information Processing Systems 15. Cambridge: MIT Press, 2002. 505−512.

[9] Davis JV, Kulis B, Jain P, Sra S, Dhillon IS. Information-Theoretic metric learning. In: Proc. of the 24th Int'l Conf. on Machine Learning. Madison: Omnipress, 2007. 209−216. [doi: 10.1145/1273496.1273523]

[10] Bilenko M, Basu S, Mooney R. Integrating constraints and metric learning in semi-supervised clustering. In: Proc. of the 21th Int'l Conf. on Machine Learning. Madison: Omnipress, 2004. 81−88. [doi: 10.1145/1015330.1015360]

[11] Kamvar SD, Klein D, Manning CD. Spectral learning. In: Proc. of the 17th Int'l Joint Conf. on Artificial Intelligence. San Francisco: Morgan Kaufmann Publishers, 2003. 561−566.

[12] Kulis B, Basu S, Dhillon I, Mooney R. Semi-Supervised graph clustering: A kernel approach. Machine Learning, 2009,74(1):1−22. [doi: 10.1007/s10994-008-5084-4]

[13] Yu SX, Shi JB. Segmentation given partial grouping constraints. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2004, 26(2):173−180. [doi: 10.1109/TPAMI.2004.1262179]

[14] Meila M, Shi JB. A random walks view of spectral segmentation. In: Proc. of the Int'l Workshop on AI and Statistics. Amsterdam: Elsevier Science & Technology Books, 2001. 873−879.

[15] De Bie T, Suykens JAK, De Moor B. Learning from general label constraints. In: Proc. of the Joint IAPR Int'l Workshops on Structural, Syntactic, and Statistical Pattern Recognition. Berlin, Heidelburg: Springer-Verlag, 2004. 671−679. [doi :10.1007/978-3-540-27868-9_73]

[16] Lu ZD, Carreira-Perpinan MA. Constrained spectral clustering through affinity propagation. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society, 2008. 1−8. [doi: 10.1109/CVPR.2008.4587451]

[17] Zhu X, Ghahramani Z, Lafferty J. Semi-Supervised learning using Gaussian fields and harmonic functions. In: Proc. of the 20th Int'l Conf. on Machine Learning. Madison: Omnipress, 2003. 912−919.

[18] Azran A. The rendezvous algorithm: Multiclass semi-supervised learning with Markov random walks. In: Proc. of the 24th Int'l Conf. on Machine Learning. Madison: Omnipress, 2007. 49−56.

[19] Shi JB, Malik J. Normalized cuts and image segmentation. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2000,22(8): 888−905. [doi: 10.1109/34.868688]

[20] Ng AY, Jordan MI, Weiss Y. On spectral clustering: Analysis and an algorithm. In: Advances in Neural Information Processing Systems. Alberta: Curran Associates Inc., 2002. 849−856.

[21] Li ZG, Liu JZ, Tang XO. Constrained clustering via spectral regularization. In: Proc. of the Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society, 2009. 421−428. [doi: 10.1109/CVPR.2009.5206852]

[22] Cai D, He X, Han J, Huang TS. Graph regularized non-negative matrix factorization for data representation. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2011,33(8):1548−1560. [doi: 10.1109/TPAMI.2010.231]

[23] Lin F, Cohen WW. Power iteration clustering. In: Proc. of the 27th Int'l Conf. on Machine Learning. Madison: Omnipress, 2010. 655−662.

附中文参考文献:

[2] 肖宇,于剑.基于近邻传播算法的半监督聚类.软件学报,2008,19(11):2803−2813. http://www.jos.org.cn/1000-9825/19/2803.htm [doi: 10.3724/SP.J.1001.2008.02803]

[3] 尹学松,胡恩良,陈松灿.基于成对约束的判别型半监督聚类分析.软件学报,2008,19(11):2791−2802. http://www.jos.org.cn/1000-9825/19/2791.htm [doi: 10.3724/SP.J.1001.2008.02791]

[4] 王红军,李志蜀,戚建淮,成飚,周鹏,周维.基于贝叶斯网络的半监督聚类集成模型.软件学报,2010,21(11):2814−2825. http://www.jos.org.cn/1000-9825/3683.htm [doi: 10.3724/SP.J.1001.2010.03683]

何萍(1983－),女,江苏太仓人,博士,讲师,主要研究领域为机器学习,数据挖掘.
E-mail: angeletx@gmail.com

陆林(1989－),男,硕士生,主要研究领域为机器学习,数据挖掘,生物信息学.
E-mail: 8987lu@sina.com

徐晓华(1979－),男,博士,副教授,主要研究领域为机器学习,数据挖掘,生物信息学,并行计算.
E-mail: arterx@gmail.com

陈崚(1951－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为人工智能,数据挖掘,系统优化.
E-mail: yzulchen@gmail.com