ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.19, No.11, November 2008, pp.2803−2813
DOI: 10.3724/SP.J.1001.2008.02803

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

基于近邻传播算法的半监督聚类*

肖 宇+, 于 剑

(北京交通大学 计算机与信息技术学院,北京 100044)

Semi-Supervised Clustering Based on Affinity Propagation Algorithm

XIAO Yu+, YU Jian

(School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China)

+ Corresponding author: E-mail: 06120567@bjtu.edu.cn

Xiao Y, Yu J. Semi-Supervised clustering based on affinity propagation algorithm. *Journal of Software*, 2008, 19(11):2803−2813. http://www.jos.org.cn/1000-9825/19/2803.htm

Abstract: A semi-supervised clustering method based on affinity propagation (AP) algorithm is proposed in this paper. AP takes as input measures of similarity between pairs of data points. AP is an efficient and fast clustering algorithm for large dataset compared with the existing clustering algorithms, such as *K*-center clustering. But for the datasets with complex cluster structures, it cannot produce good clustering results. It can improve the clustering performance of AP by using the priori known labeled data or pairwise constraints to adjust the similarity matrix. Experimental results show that such method indeed reaches its goal for complex datasets, and this method outperforms the comparative methods when there are a large number of pairwise constraints.

Key words: semi-supervised clustering; affinity propagation; similarity matrix; pairwise constraints; prior knowledge

摘 要: 提出了一种基于近邻传播(affinity propagation,简称 AP)算法的半监督聚类方法.AP 是在数据点的相似度矩阵的基础上进行聚类.对于规模很大的数据集,AP 算法是一种快速、有效的聚类方法,这是其他传统的聚类算法所不能及的,比如:*K* 中心聚类算法.但是,对于一些聚类结构比较复杂的数据集,AP 算法往往不能得到很好的聚类结果.使用已知的标签数据或者成对点约束对数据形成的相似度矩阵进行调整,进而达到提高 AP 算法的聚类性能.实验结果表明,该方法不仅提高了 AP 对复杂数据的聚类结果,而且在约束对数量较多时,该方法要优于相关比对算法.

关键词: 半监督聚类;近邻传播;相似度矩阵;成对点约束;先验知识

中图法分类号: TP181 **文献标识码**: A

* Supported by the National Natural Science Foundation of China under Grant No.60875031 (国家自然科学基金); the National Basic Research Program of China under Grant No.2007CB311002 (国家重点基础研究发展计划(973)); the Program for New Century Excellent Talents in University of China under Grant No.NECT-06-0078 (新世纪优秀人才支持计划); the Research Fund for the Doctoral Program of Higher Education of the Ministry of Education of China under Grant No.20050004008 (教育部高等学校博士学科点专项科研基金); the Fok Ying-Tong Education Foundation for Young Teachers in the Higher Education Institutions of China under Grant No.101068 (霍英东教育基金会高等院校青年教师基金)

Received 2008-03-01; Accepted 2008-08-26

聚类算法是一种有效的数据分析方法,聚类算法是在没有任何数据的先验信息下对数据进行聚类分析的,这类算法又称为无监督学习方法.在很多实际问题中,由于对数据本身没有任何先验分析,这种传统的聚类算法有时得不到有效的聚类结果.在实际问题中,有时我们会获得少部分数据的先验知识,包括类标签和数据点的划分约束条件(比如成对约束信息).如何利用这些仅有的少量先验知识来对大量没有先验知识的数据进行聚类分析成为一个非常重要的问题.半监督聚类则是针对这类问题提出的,即利用少量具有先验知识的数据来辅助无监督聚类,所以半监督聚类逐渐成为聚类分析研究的热门问题[1−14].除了半监督聚类算法之外,一些文献中提出了 supervised clustering 算法[15,16],在此称为监督聚类算法.监督聚类算法是假设大部分数据具有先验知识,即已被正确划分,算法目的是使得到的聚类结果尽量与已知的先验知识保持一致.与半监督聚类算法相比,监督聚类算法需要更多的先验知识来保证聚类结果的有效性.这导致了监督聚类算法的目标函数和半监督聚类算法有很大的不同.一般说来,监督聚类的目标函数由两部分组成,包括类不纯度指标和聚类簇的数目.类不纯度表示聚类簇中包含的非主流标签数据所占的比例,所谓非主流标签数据是指这样一些数据,其具有的类标签不是该聚类簇中出现的最频繁的类标签,显然,类不纯度越低越好.此外,算法倾向于得到聚类数较小的聚类结果.与半监督聚类相比,监督聚类需要提供的先验信息必须是类标签,成对约束信息不能用于监督聚类算法.从上述分析可知,半监督聚类比监督聚类需要的先验信息要少许多,也要弱许多.

半监督聚类算法大致可以分为两类.一类是基于约束的(constraint-based)半监督聚类算法.这类算法是利用标签数据或者成对约束信息来改进聚类算法本身.常用的方法有:(1) 通过修改聚类的目标函数来满足成对约束.如,文献[1]是通过在原始目标函数中添加基于类标签数据的类不纯度指标来约束聚类过程,目的是使具有相同类标签的数据能被划分在同一聚类簇中,文献[2,3]是通过在原始目标函数中添加基于约束对信息的惩罚项来对违反先验约束对信息的数据划分进行惩罚,其目的是尽可能地得到满足给定约束对信息的划分.(2) 在聚类过程中遵循约束条件,使得到的聚类结果满足所有约束对信息.如,Wagstaff 在文献[4]中提出的 COP-COWEB 算法和文献[5]中提出的 COP-Kmeans 算法要求每一步划分都满足已知的约束对信息,最终得到满足所有约束对信息的聚类结果.(3) 依照标签数据初始化聚类参数并约束聚类过程.如,文献[2]利用约束对先验知识来指导初始聚类中心的选取,使初始类中心满足已知的约束条件,文献[6]是基于类标签数据进行初始类中心选择的,基本思想是将具有相同类标签的标签数据看成一个子集,假设数据划分为 K 类,这样可以得到 K 个不相交的子标签数据集,K 个子数据集的质心作为 K 个初始类中心点.另一类是基于距离的(metric-based 或者 distance-based)半监督聚类算法.这类算法是利用标签数据或者成对约束信息学习一种新的距离测度函数来满足约束条件.常用的方法有:(1) 利用成对约束信息和最短路径算法来调整距离测度.如,文献[7,8]利用成对约束信息来调整距离矩阵,文献[9,10]利用最短路径算法对约束信息调整的距离矩阵再进行调整,目的是使距离矩阵能够更充分地反映已知的约束对信息.(2) 利用约束对信息构建最优化问题,通过求解此凸优化问题来得到新的距离测度函数[11,12].(3) 利用同类约束对信息来学习新的马氏距离矩阵,利用新得到的距离进行聚类[13].(4) 利用成对约束信息对原始数据进行基于约束的特征投影,在得到的新的子空间进行聚类[14].除了这两类基本的半监督聚类方法以外,上述的一些算法是结合这两种基本思想得到的半监督聚类算法[2,3,14].

本文提出的基于近邻传播(affinity propagation,简称 AP)算法的半监督聚类算法也可以被当作一种基于距离测度函数学习的方法.近邻传播聚类(affinity propagation clustering,简称 APC)[17,18]是由《Science》中的一篇文章提出来的.与以往的聚类方法相比,此方法可以更快地处理大规模数据,得到较好的聚类结果.文献中将近邻传播聚类应用在人脸图像聚类、基因表达数据的基因识别、手写体字符识别、最优航空路线确定等问题上,实验结果表明,近邻传播聚类在很短的时间内就能得到 K 中心算法花费很长时间才能达到的聚类结果.近邻传播聚类的另一个优点是,它对数据形成的相似矩阵的对称性没有任何要求,这样也就扩大了它的应用范围.但是,对于一些本身具有复杂结构的数据集,近邻传播聚类通常不能得到合理的聚类结果.本文将原始近邻传播算法与半监督思想相结合,启发式地引入已知的标签数据或成对点约束来调整相似度矩阵.通过在新得到的相似度矩阵的基础上进行近邻传播聚类,达到提高聚类性能的目的.实验结果表明,半监督的近邻传播算法性能与原始聚类算法性能相比有明显的提高.对比实验结果也表明,从所有的实验数据集的整个实验结果来分析,半监督

近邻传播算法具有一定的优势.

1 近邻传播聚类

近邻传播聚类(AP)算法是一种基于近邻信息传播的聚类算法,其目的是找到最优的类代表点集合(一个类代表点对应为实际数据集中的一个数据点,exemplar),使得所有数据点到最近的类代表点的相似度之和最大.如果设数据点的相似度为数据点的欧式距离的负数,则 AP 算法的目标函数与经典的 K 中心聚类(K-center clustering)算法的目标函数一致,但是其算法原理与 K 中心算法的原理存在很大的不同.AP 算法将每个数据点看成图中的一个节点,通过在图中进行信息传播来找到最优的类代表点集合.K 中心算法迭代过程则是通过不断更新聚类中心来提高聚类质量.AP 算法是基于数据点的相似度信息进行传播得到最优类代表点来优化目标函数,K 中心算法则是基于代价最小替换原则得到最优类中心来优化目标划分准则.此外,AP 算法与 K 中心算法采用了不同的方法来确定初始类代表点,AP 算法将所有数据点都作为候选的类代表点,这样就避免了聚类结果受限于初始类代表点的选择.K 中心算法则是随机选择几个点作为初始类代表点,致使聚类结果对初始类代表点的选择非常敏感.AP 算法与一般聚类算法相比,最大的优点在于,AP 算法对相似度矩阵的对称性没有要求,这也就扩大了 AP 算法的应用范围.AP 算法之所以称为近邻传播算法在于近邻点的信息直接影响了算法中信息的传播结果,下面的公式(4)~公式(7)给出了详细的解释.从迭代公式(4)~公式(7)中可以看出,每一次信息传播都是由数据点与最近邻点或次近邻点的信息计算得到的.

AP 算法是在数据形成的相似度矩阵的基础上进行聚类的,本文选用欧式距离作为相似度的测度指标.求解任意两点之间的相似度为两点距离平方的负数,例如,对于点 x_i 和点 x_k 则有 $s(i,k)=-\|x_i-x_k\|^2$. AP 方法用 $s(i,k)$ 表示数据点 x_k 在多大程度上适合作为数据点 x_i 的类代表点.AP 算法要为每个数据点 k 设定其偏向参数 $s(k,k)$(*preference*)的值,$s(k,k)$的值越大,相应的点 k 被选中作为类代表点的可能性也就越大.AP 算法初始假设所有数据点被选中成为类代表点的可能性相同,即设定所有 $s(k,k)$为相同值 p.同样,p 值的大小也影响到最终得到聚类的类的个数,AP 算法可以通过改变 p 值来寻找合适的类的数目(实验结果说明,一般情况下,增大 p 值可以增加类的个数,减小 p 值可以减少类的个数),这是 AP 算法中的一个重要参数.

AP 算法引入了两个重要的信息量参数,分别定义为代表矩阵 $R=[r(i,k)]_{n\times n}$ 和适选矩阵 $A=[a(i,k)]_{n\times n}$. AP 算法的迭代过程就是这两个信息量交替更新的过程,两个信息量代表了不同的竞争目的.$r(i,k)$(*responsibility*)是从点 x_i 指向点 x_k,它代表点 x_k 积累的证据,用来表示 x_k 适合作为 x_i 的类代表点的代表程度.$a(i,k)$(*availability*)是从点 x_k 指向点 x_i,它代表点 x_i 积累的证据,用来表示 x_i 选择 x_k 作为类代表点的合适程度.对于任意数据点 x_i,计算所有数据点的代表程度 $r(i,k)$和适选程度 $a(i,k)$之和,则 x_i 的类代表点为 x_k: $\arg\max_k(a(i,k)+r(i,k))$.AP 算法的核心步骤为两个信息量的交替更新过程,更新公式如下:

$$r(i,k)\leftarrow s(i,k)-\max_{k's.t.k'\neq k}\{a(i,k')+s(i,k')\} \tag{1}$$

$$\text{If}\quad i\neq k,\ a(i,k)\leftarrow\min\left\{0,r(k,k)+\sum_{i's.t.i'\notin\{i,k\}}\max\{0,r(i',k)\}\right\} \tag{2}$$

$$a(k,k)\leftarrow\sum_{i's.t.i'\neq k}\max(0,r(i',k)) \tag{3}$$

对上述迭代公式(1)两边同时加上 $a(i,k)$ 得到如下变形:

$$r(i,k)+a(i,k)\leftarrow s(i,k)+a(i,k)-\max_{k's.t.k'\neq k}\{a(i,k')+s(i,k')\} \tag{4}$$

为了更好地解释公式(4)代表的意义,在此给出一些定义: $E=[e(i,k)]_{n\times n}=R+A$ 称为决策矩阵, $\Gamma=[\tau(i,k)]_{n\times n}=s+A$ 称为潜力矩阵.在数据集 X 中,数据点 i 关于相似矩阵 s 的最近邻表示为 i_*^s,在此, $i_*^s=\arg\max_{k\in\{1,2,\ldots,n\}\setminus\{i\}}\{s(i,k)\}$. 数据点关于相似度矩阵 s 的次近邻表示为 i_{**}^s,这里, $i_{**}^s=\arg\max_{k\in\{1,2,\ldots,n\}\setminus\{i,i_*^s\}}\{s(i,k)\}$.基于上述定义,公式(4)可以简单地表示为 $e(i,k)\leftarrow\tau(i,k)-\max_{k's.t.k'\neq k}\{\tau(i,k')\}$.

当 AP 算法到达收敛时,决策矩阵 E 有以下 3 种表达形式:

当 $k=i_*^\Gamma$ 时,

$$e(i,k)=\tau(i,i_*^\Gamma)-\max\left(\tau(i,i_{**}^\Gamma),\tau(i,i)\right) \tag{5}$$

当 $k\neq i_*^\Gamma, i$ 时,

$$e(i,k)=\tau(i,k)-\max\left(\tau(i,i_*^\Gamma),\tau(i,i)\right) \tag{6}$$

当 $k=i$ 时,

$$e(i,i)=\tau(i,i)-\tau(i,i_*^\Gamma) \tag{7}$$

从公式(5)~公式(7)可以很明显地看出,决策矩阵中元素值的大小直接受到数据点关于潜力矩阵的最近邻或次近邻的信息量的影响,这充分体现了近邻信息在 AP 算法中起到了决定性作用,所以,本文称 AP 算法为近邻传播算法.

AP 算法在信息更新这一步骤引入了另外一个重要的参数 λ,称为阻尼因子(damping factor).在每一次循环迭代中,$r(i,k)$和 $a(i,k)$的更新结果都是由当前迭代过程中更新的值和上一步迭代的结果加权得到的.设当前迭代次数为 t,加权公式为

$$r^{(t)}(i,k)=(1-\lambda)\times\left(s(i,k)-\max_{k's.t.k'\neq k}\left\{a^{(t-1)}(i,k)+s(i,k')\right\}\right)+\lambda\times r^{(t-1)}(i,k).$$

$$\text{If}\quad i\neq k,\ a^{(t)}(i,k)=(1-\lambda)\times\left(\min\left\{0,r^{(t-1)}(k,k)+\sum_{i's.t.i'\neq\{i,k\}}\max\left\{0,r^{(t)}(i',k)\right\}\right\}\right)+\lambda\times a^{(t-1)}(i,k).$$

$$a^{(t)}(k,k)=(1-\lambda)\times\left(\sum_{i's.t.i'\neq\{i,k\}}\max\left\{0,r^{(t)}(i',k)\right\}\right)+\lambda\times a^{(t-1)}(i,k).$$

其中,$\lambda\in[0,1)$,默认值为 0.5[17],阻尼因子的作用是避免 AP 算法发生振荡,增大阻尼因子可以消除振荡.在本文的实验中,为了避免振荡的发生,设置阻尼因子为 0.9.下面给出 AP 算法程序的基本步骤.

(1) 初始化:求解相似度矩阵 $[s(i,k)]_{n\times n}$,n 为数据点的个数,设定相似度矩阵对角线元素 $s(k,k)$为一相同值 p<0.初始化 *availabilities* 和 *responsibilities* 为 0:$a^{(0)}(i,k)=r^{(0)}(i,k)=0$.

(2) 迭代过程:

① 信息量 *responsibilities* 和 *availabilities* 根据公式(1)~公式(3)和加权公式进行更新.

② 对所有数据点求和信息量 *responsibilities* 和 *availabilities*,找到每个点的类中心点.

③ 判断信息迭代过程是否满足停止条件:超过某一迭代最大数目;信息改变量低于某一固定阈值;选择的类中心在连续几步迭代过程中保持稳定.满足其中一个停止条件即可.

(3) 判断得到的类中心的个数是否满足要求,如果不满足,则改变 p 值,重复进行程序直至聚类个数满足要求为止,输出最终聚类结果.

2 基于AP算法的半监督聚类

AP 算法中最重要的两个参数为相似度矩阵 s 和偏向参数 p,如果相似度矩阵能够准确地给出数据之间的相似关系,则 AP 算法就能得到很好的聚类结果.相似度矩阵的定义直接影响到基于相似度矩阵的聚类算法的性能.偏向参数的大小可以改变聚类数的多少,它作为数据点独立的信息,只反映了每个数据点被选中为代表点的概率的大小,所以,利用先验信息来辅助偏向参数的确定的可操作性不强.但是,由于相似度矩阵包含了数据对之间的信息,因此,利用先验的数据成对约束调整相似矩阵更为合理,也更容易.注意到 AP 算法已经由实验证实比 *K*-center 性能要好[17],因此,本文提出了基于相似度矩阵调整的半监督的 AP 算法,即 SAP(semi-supervised affinity propagation)算法.SAP 算法是利用标签的数据信息来调整点与点之间的相似度形成新的相似度矩阵 s,在新得到的相似度矩阵的基础上进行 AP 算法.在半监督聚类中,我们得到的数据信息一般是带类标签的数据或者是成对点约束信息.在很多实际问题中,获得成对约束信息相对于获得类标签信息要容易些,而且类标签信

息可以转化为成对点约束信息,反之则不然.所以,大部分半监督聚类算法是基于成对约束信息提出来的,SAP算法也是利用成对约束作为先验信息.成对点约束分为两种[4]:Must-link,两个点必须属于同一类,即集合$M=\{(x_i,x_j)\}$;Cannot-link,限制规定两个点不能在同一类中,即集合$C=\{(x_i,x_j)\}$.这里,我们使用的成对点约束信息进行实验,可将获得的带类标签的数据转化为成对点约束来进行计算.

基于AP算法的半监督聚类的核心步骤是相似度矩阵的调整.本文对相似度矩阵s的调整所依据的基本原则是:当约束对$\{(x_i,x_j)\}\in M$时,即认为两数据点具有很高的相似性,从第1节给出的相似性度量可知相似度最高为0,所以将同类约束对数据之间的相似度调整为$s(i,j)=0$;当约束对$\{(x_i,x_j)\}\in C$时,即认为两数据相似度很低,相似性度量中定义最低相似度为$-\infty$,所以将不同类约束对数据间的相似度调整为$s(i,j)=-\infty$.

除了对已知的约束对数据间的相似度进行调整以外,还参考文献[2,9]中的距离矩阵调整方法对其他点之间的相似度进行调整.通过进一步的调整可以得到更多的同类约束对信息和不同类约束对信息,这增加了约束对信息的获得量.此外,一些非约束对数据点的相似度由于约束对数据之间的相似度的改变也发生变化,对于这些数据的相似度也进行了调整.相似度矩阵S的调整步骤如下:

(1) 对先验信息中满足must-link约束的数据对以及由初始的must-link约束集根据must-link关系的传递性扩展得到的新的满足must-link约束的数据对的相似度进行调整.对于先验信息中已有的must-link约束对,

$$(x_i,x_j)\in M\Rightarrow s(i,j)=0\,\&\,s(j,i)=0.$$

由已知的must-link约束对扩展得到新的must-link约束对,将新的must-link约束加入到must-link约束集中:

$$(x_i,x_k)\notin M\,\&\,(x_i,x_j)\in M\,\&\,(x_j,x_k)\in M\Rightarrow s(i,k)=0\,\&\,s(k,i)=0\,\&\,M=(x_i,x_k)\cup M.$$

通过这一调整,增加了must-link约束对的数量,这也对后面的调整步骤(3)~步骤(4)有直接影响.

(2) 对先验信息中满足cannot-link约束的数据对的相似度进行调整:

$$(x_i,x_j)\in C\Rightarrow s(i,j)=-\infty\,\&\,s(j,i)=-\infty.$$

(3) 基于上述两步的初步调整结果,基于最短路径原则(这里为最大相似度原则)对不包含在先验信息中的数据对的相似度进行全局调整.如果数据集中存在一个数据点与待调整的数据对分别相连,这一数据点与这对数据点的相似度之和大于这对数据点的初始相似度,则调整这对数据点的相似度为较大的相似度.公式化为

$$(x_i,x_j)\notin\{M\cup C\}\Rightarrow s(i,j)=\max(s(i,j),s(i,k)+s(k,j)).$$

此外,考虑到只有在must-link集合中的数据点x_k才有可能导致$s(i,j)$的值发生变化,所以,计算时只与集合M中某约束对中对应的数据点进行比较来加快运算速度.通过这一调整增加了原有先验信息中的must-link约束集.

(4) 基于cannot-link约束集对第(3)步中的调整结果进行局部修正.当全局调整中选择的数据点x_k与这对数据点中的一个数据点之间满足cannot-link约束,和另一个数据点之间满足must-link约束,则认为这对数据点也满足cannot-link约束,即数据点的相似度调整为最小.公式化为

$$(x_i,x_j)\notin\{M\cup C\}\,\&\,(x_i,x_k)\in C\,\&\,(x_k,x_j)\in M\Rightarrow s(i,j)=-\infty\,\&\,s(j,i)=-\infty$$

或者

$$(x_i,x_j)\notin\{M\cup C\}\,\&\,(x_i,x_k)\in M\,\&\,(x_k,x_j)\in C\Rightarrow s(i,j)=-\infty\,\&\,s(j,i)=-\infty.$$

调整后将这一数据对加入到cannot-link集合中:$C=(x_i,x_j)\cup C$.

经过上述调整,数据的相似度矩阵变化很大.AP算法是在相似度矩阵的基础上进行信息迭代的,这就决定了整个迭代过程也将随之发生改变.从第1节给出的公式(1)~公式(3)可以看出,相似度矩阵的调整直接改变了代表程度值这一信息量的迭代过程,使得同类约束对数据间的代表程度值变高,使得不同类约束对数据间的代表程度值变为$-\infty$,这表示,算法使得同类的约束对数据点尽可能地被划分到同一类中,而不同类约束对数据点最终不能被划分到同一类中.这一改变是相似度矩阵的调整带来的直接影响.此外,从公式(2)可知,代表程度值的调整也将导致适选值计算迭代过程发生改变,公式(1)又表明代表程度值的变化同样受到适选值的影响.这一循环迭代过程表明了相似度的改变将最终导致所有数据的代表程度值和适选值都将发生改变,也就导致了算法最终收敛的结果发生很大的变化.AP算法通过近邻信息的传播将约束对信息产生的影响也加以传播,使得最终的聚类结果发生变化.

基于上述调整得到新的相似度矩阵之后,利用AP算法基于新的相似度矩阵进行聚类.因为先验知识只是

改变了局部数据的相似度,SAP 并不能保证聚类结果满足先验知识给定的所有约束对信息,即存在两种违反约束对信息的情形:给定的属于同一类的约束对数据被分到了不同的类中,给定的属于不同类的数据被分到了同一类中.为了解决这一问题,本文在 AP 算法的聚类结果的基础上对违反的约束对信息的类别归属进行如下调整:

首先,给定数据集 X 被分为 K 类,得到 K 个类代表点:$\{x_1,x_2,...,x_K\}$.用 $y_i=k(k=1,\ldots,K)$表示 x_i 类标签,k 对应为类代表点 x_k,即 $y_i=k$ 表示 x_i 选择 x_k 作为其类代表点.假定聚类结果得到的类代表点符合数据集的内部结构,即得到的类代表点分别分布在数据集的 K 个子结构中,则认为数据点离各自类代表点越近,被划分正确的概率越大.下面的调整是基于上述假设进行的.

(1) 对违反 must-link 约束对数据的调整.已知$(x_i,x_j)\in M, y_i=k, y_j=k'$,分别计算两个数据点到这两个类代表点的距离之和,即 $d_{ik}+d_{jk}$ 和 $d_{ik'}+d_{jk'}$.如果 $(d_{ik}+d_{jk})<(d_{ik'}+d_{jk'})$,则改变 x_j 的类标签为 $y_i=k$;否则,改变 x_i 的类标签为 $y_i=k'$.

(2) 对违反 cannot-link 约束对数据的调整.已知 $(x_i,x_j)\in C, y_i=y_j=k$,分别计算两个数据点到各自类代表点的距离 d_{ik} 和 d_{jk}.如果 $d_{ik}<d_{jk}$,则保持 x_i 的类标签不变,改变 x_i 的类标签为 $y_i=\arg\min_{k',k'\neq k}(d_{jk'})$.

3 实 验

本文选择 6 个数据集对 SAP 算法和相关聚类算法进行比对实验.下面分别对实验数据、比对算法以及实验结果进行介绍.

3.1 实验数据

实验中用到的数据集为 UCI 数据库中的数据集.其中,标准数据集有 Iris 数据集、Ionosphere 数据集、Glass 数据集;从手写字母识别数据库中随机抽取的 3 类字母:{I,J,L};从手写数字识别数据库中随机抽取的数字集:{3,8,9},{1,2,3,4},抽样率都为 10%.表 1 中给出了数据集的有关信息.

Table 1 Datasets used in experiment

表 1 实验中使用的数据集

	Iris	Ionosphere	Glass	Letter{I,J,L}	Digits-1{3,8,9}	Digits-2{1,2,3,4}
Instance	150	351	214	227	317	448
Dimension	4	34	9	16	16	16
Class	3	2	6	3	3	4

在实验中,对于每个数据集,我们采用 5 重交叉检验成对约束对聚类结果的影响.每次从原始数据集中抽取 80%作为训练数据集,剩余的 20%作为测试数据集.成对约束是从训练数据集中随机产生的.用随机产生的约束对来指导聚类算法对全部数据的聚类分析,最后用聚类评价指标对测试数据集的聚类结果进行评价.在每一确定数量的成对约束下,对每一数据集重复实验 20 次(5 重交叉验证/次),取 20 次 5 重交叉验证评价结果的均值代表半监督聚类算法在固定数量的成对约束条件下对某一数据集的聚类性能.每次只对测试数据集进行聚类结果的评估,目的在于避免训练数据集中已知的约束对信息对聚类性能评价结果的影响.

3.2 比对算法

SAP 算法是利用先验知识调整相似矩阵来改善 AP 算法的聚类结果.SAP 是基于距离测度学习的半监督聚类算法的一种.在此,我们选择几种有效的基于距离测度学习的半监督聚类算法进行比对实验.实验采用的比对算法大致可以概括为两类:一类是结合 K-means 算法的半监督聚类算法,另一类是在近来比较流行的谱聚类算法的基础上提出的半监督聚类算法.K-means 算法和谱聚类算法采用两种截然不同的聚类方法对数据进行处理.K-means 对于处理球形数据可以给出很好的聚类结果,但在分析具有非规则形状类的数据集时得不到有效的聚类结果.谱聚类成为近年来比较流行的聚类算法的原因之一在于它可以较好地解决对包含不规则形状类的数据集的聚类分析.

3.2.1 基于 *K*-means 的半监督聚类算法

Wagstaff 等人最早将成对约束信息用于 *K*-means 算法,提出了 COP-Kmeans(constraint-partitioning-*K*-means)算法[5].Klein 等人提出了基于约束的层次聚类[9],通过约束信息和最短路径算法来调整距离矩阵,利用层次聚类算法对调整后的数据进行聚类.随后,Hillel 等人提出了基于 RCA(relevant component analysis)算法[12]即相关成分分析算法的半监督学习方法,此算法利用约束对信息求解新的距离测度得到马氏距离,利用新的测度函数将原始数据集进行投影得到新的数据集,最后用传统的聚类算法对新的数据集进行聚类.其实验结果表明,此算法的聚类性能优于 Xing 等人[11]提出的距离测度学习算法.本文选用 RCA 与 *K*-means 算法结合与 SAP 算法进行比对.

SCREEN(semi-supervised clustering method based on spherical *K*-means via feature projection)算法[14]是最近由 Tang 等人提出的一种通过特征投影的约束的球形 *K*-means 半监督聚类算法.该算法首先利用成对约束对原始数据进行基于约束的特征投影来对数据降维,然后在得到的特征子空间上对数据用基于约束的球形 *K*-means 进行聚类.基于约束的球形 *K*-means 算法与前面提到的约束的 *K*-means 算法不同,它采用一种启发式的解决方案来放松地执行成对约束条件.RCA 算法与 SCREEN 可以用来降维.本文的实验中设置特征投影后的数据集与原始数据集保持相同的维数.

3.2.2 基于谱聚类的半监督聚类算法

谱聚类算法是一种基于图划分理论的方法,它是将图划分问题放松为求解图的 Laplacian 矩阵的谱分解问题.谱聚类方法最早是由 Shi 和 Malik[19]提出来的,将规范割方法用于解决图像分割问题.谱聚类在数据相似度矩阵的基础上进行,在得到的相似度矩阵的基础上求解 Laplacian 矩阵(Laplacian 矩阵有多种求解方法,不同的谱聚类算法求解方法可能不同).相似性度量的定义与聚类的性能有直接的关系,半监督的谱聚类就是利用成对约束信息对数据之间的相似性进行调整来达到更好的聚类结果.

Kamvar 等人[7]根据成对约束对相似性矩阵进行调整,它只是将已知的 must-link 数据对的相似性设置为 1,cannot-link 数据对的相似性设置为 0,而没有在此基础上对其他数据点之间的相似性进行调整.同样,Xu 等人[8]提出的 CSC(constrained spectral clustering)算法也采用这样的方法对相似性矩阵进行调整.王玲等人提出了一种密度敏感的半监督谱聚类算法:DS-SSC(density-sensitive semi-supervised spectral clustering)算法[11].与前面两种算法不同,DS-SSC 算法首先根据成对约束信息调整距离矩阵,然后利用密度敏感的距离测度将约束信息进行空间传播.该算法不仅利用了成对约束先验信息,而且还考虑了数据的空间一致性结构信息,间接地调整了相似性矩阵.其实验结果表明,DS-SSC 算法的聚类性能要优于仅考虑成对约束信息的 CSC 算法.

实验中采用 DS-SSC 算法和 CSC 算法与 SAP 算法进行比对实验.谱聚类和 AP 算法都是基于数据的相似矩阵进行聚类的.AP 算法比谱聚类算法适用范围要广,因为 AP 算法对数据的相似矩阵的对称性没有任何要求,这是 AP 算法的一大优点.

3.3 实验结果及分析

实验中采用两种不同的评价指标对聚类结果进行评价:成对 *F*-评测指标和 CRI 指标.成对 *F*-评测指标是基于传统的信息重获评测指标提出的,主要是针对同属于一类的成对点的预测情况来判定聚类结果的质量.成对 *F* 指标可以测试部分数据集的聚类结果,分别对五重交叉验证中测试数据集的聚类结果给出评价.成对 *F* 指标是由准确率(precision)和召回率(recall)两个指标组合形成的,其定义如下:

$$Precison=\frac{Pairs\ Correctly\ Predicted\ In\ Same\ Cluster}{Total\ Pairs\ Predicted\ In\ Same\ Cluster} \tag{8}$$

$$Recall=\frac{Pairs\ Correctly\ Predicted\ In\ Same\ Cluster}{Total\ Pairs\ In\ Same\ Cluster} \tag{9}$$

$$F=\frac{2\times Precision\times Recall}{Precision+Recall} \tag{10}$$

CRI 指标是对 Rand 指标的改进.Rand 指标是常用的聚类性能的评价指标,是对聚类结果中每一成对点是否

来自同一类进行判断,定义如下: $RI = correct\ decisions / total\ decisions$. 为了将 Rand 指标用于对半监督聚类算法的评价,Klein 等人对 Rand 指标进行修正,得到 CRI 指标:

$$CRI = \frac{correct\ \ free\ decisions}{total\ \ free\ decisions} \tag{11}$$

在此,$total\ \ free\ decisions = (n \times (n-1))/2 - Cn$, n 为数据点的数目,Cn 表示约束对的数目.$correct\ free\ decisions$ 为划分正确的数据对的数目减去约束对中划分正确的数据对的数目.

从上述两个评价指标的定义可知,F 指标是针对同类数据点对的聚类结果进行评价,着重于对同类数据点对的聚类评价;CRI 指标是对数据点是否来自同一类进行判断,指的是通过判断不同类之间的数据对是否被划分在不同类中,同一类的数据对是否仍被划分在同一类中对聚类结果进行评价.实验中,我们利用 F 指标对部分数据测试数据集进行评价,利用 CRI 指标对整个数据集进行评价.通过比较两个评价指标得到的评价结果更好地对聚类算法进行评价.

实验对 SAP 算法、DS-SSC 算法、CSC 算法、SCREEN 算法以及 RCA+K-means 算法分别在 6 个数据集上进行测试.实验结果如图 1 所示.

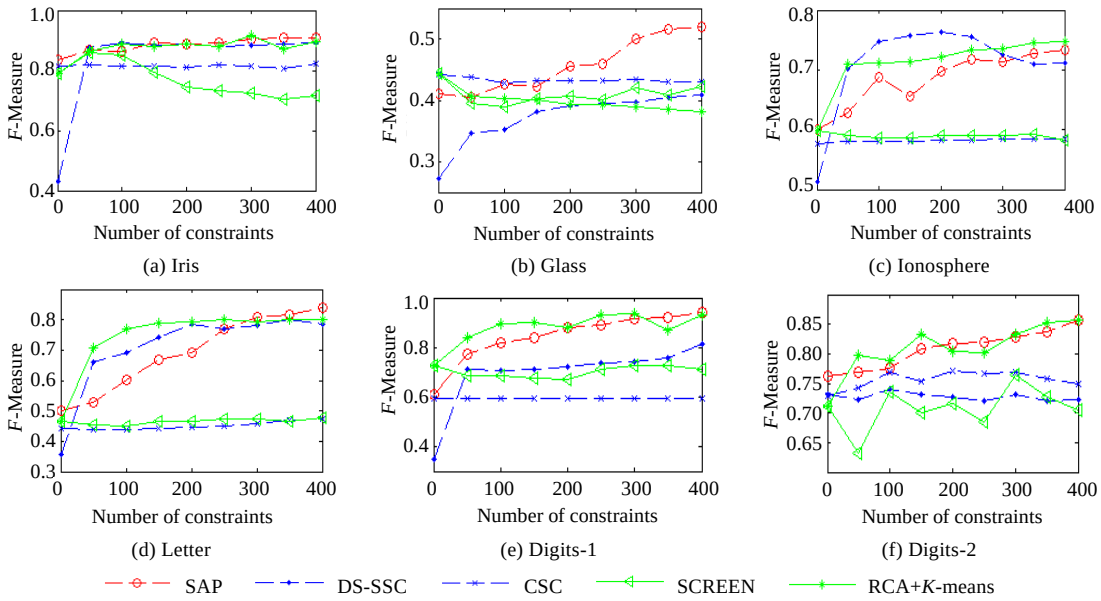


Fig.1 Clustering performance evaluated by F index

图 1 评价指标 F 对聚类算法性能的评价结果

图 1 给出了 6 个数据集上不同聚类算法在不同数量的约束对下的 F 指标值.其中,虚线表示对相似度矩阵作调整的半监督聚类算法,实线表示结合 K-means 算法的半监督聚类算法.在约束对为 0 时,CRI 指标为 AP 算法的聚类性能,从图中可以看出,SAP 算法与 AP 算法相比,其聚类性能得到了很大的提高.除了数据集 Ionospher 以外,SAP 算法在其他几个数据集上给出了较好的聚类结果.对于所有数据集,SAP 算法随着约束对数目的增加,其聚类性能都在逐渐得到改善.从图中可以看出,RCA+K-means 在大部分数据集上也得到了较好的聚类结果,但从图 1(b)可以看出,对于数据集 Glass,RCA+K-means 算法随着约束对数目的增加,其聚类性能明显呈现下降趋势,这说明在 RCA+K-means 算法中,约束对信息不但对数据集划分起不到任何指导作用,反而比原始 K-means 的聚类性能还差一些.对于 DS-SSC 算法,只有在数据集 Ionosphere 上聚类性能高于 SAP 算法的性能,但是从图 1(c)可以看出,DS-SSC 算法在约束对数目大于 200 时聚类性能呈现出下降趋势.其他两种比对算法在所有数据集上的 F 指标值说明了这两种算法的性能要低于其他几种算法.

从图 2 可以看出,相对于比对算法而言,SAP 算法在这 6 个数据集上都给出了较好的聚类结果.SAP 算法聚类结果随着约束对数量的增加呈现上升趋势.从图 2(f)可以看出,DS-SSC 算法的聚类性能低于无监督条件下的聚类性能.对于 RCA+K-means 算法,除了数据集 Glass 和数据集 Ionoshpere 以外,其他数据集上也给出了较好的聚类结果.CRI 指标的评价结果表明,SAP 算法通过增加少量约束对信息能够改善整个数据集的聚类结果,从图 2 中 6 个数据集的评价结果可以看出,SAP 算法相对于其他比对算法具有一定的优越性.

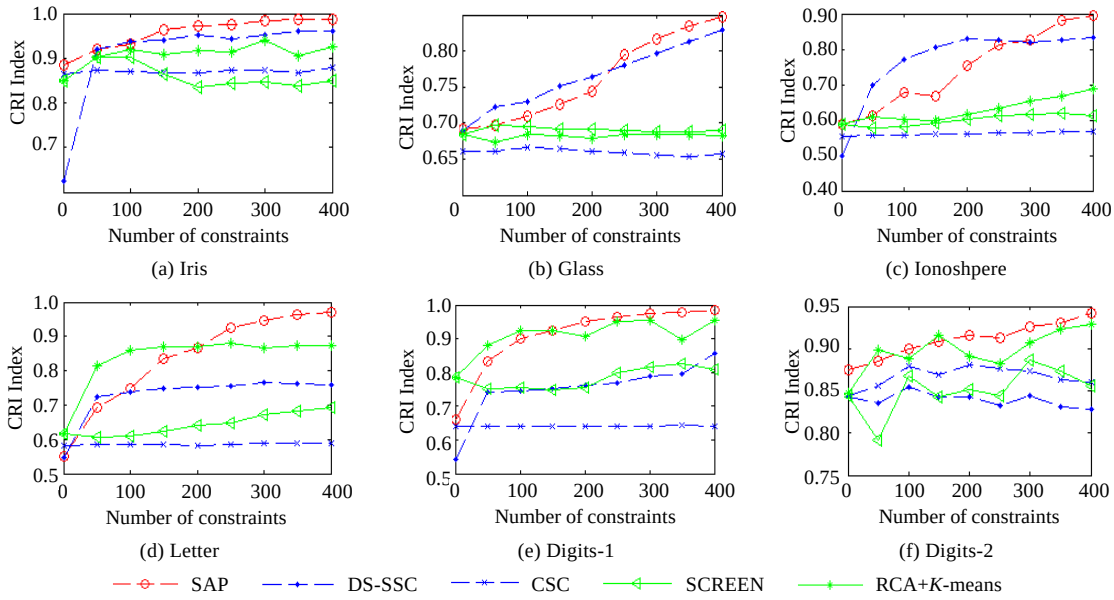


Fig.2 Clustering performance evaluated by CRI index

图 2 评价指标 CRI 对聚类算法性能的评价结果

综合两个指标的评价结果,SAP 算法都给出了较好的聚类结果.对于两个评价指标,随着约束对数量的增加,SAP 算法的聚类性能呈上升趋势,且在约束对信息比较充分的条件下,SAP 算法能够给出优于其他比对算法的聚类结果.虽然 SAP 算法只是利用约束对信息对相似度矩阵进行调整,但在相对简单的调整下能够给出较好的聚类结果不仅取决于 AP 算法本身,而且表明了相似度调整对于 AP 算法性能的提升有较大的帮助.SAP 中的相似度矩阵的调整方法与比对算法中利用成对约束学习新的测度距离的方法相比稳定性较好,因为本文给出的调整方法不会影响到数据集的结构,能够使数据集的结构变得明了.RCA 算法在大多数测试数据集上给出了合理的聚类结果,但是对于一些数据集也可能会给出不及无监督算法的聚类结果,与 SAP 算法相比,RCA 算法对数据集比较敏感.基于上述分析,从整体上说,SAP 算法的性能要优于本文实验部分的比对算法.

4 结 论

本文提出了一种基于近邻传播算法的半监督聚类方法.该方法是一种基于距离测度学习的半监督聚类算法.本文针对现有聚类技术的不足,提出在近邻传播算法的基础上,启发式地引入已知的标签数据或成对点约束来调整相似度矩阵.通过在新得到的相似度矩阵的基础上进行近邻传播聚类算法.原始的近邻传播算法处理大数据集与含有多类数据的数据集具有速度快、效果好等优势,但是对于一些复杂数据结构数据集则通常不能给出令人满意的聚类结果.SAP 算法是在近邻传播算法的基础上通过改变相似度矩阵对聚类进行指导.实验结果表明,半监督的 AP 算法在聚类性能上有了显著的提高.从实验结果可以看出,当获得较为充分的先验信息时,SAP 算法的表现要好于比对算法.SAP 算法除了在实验中测试数据集上表现了较好的聚类性能以外,由于 AP 算法对相似度矩阵的对称性没有要求,这就放宽了对数据集本身的要求,SAP 算法也因此具有相对广泛的应用

范围.

当然,SAP算法不仅继承了AP算法的优点,而且将AP算法中的参数选择问题延续下来.如何选择合适的参数以达到更好的聚类性能是任何含参数算法所亟待解决的问题.需要考虑能否在参数设定和数据集自身之间建立联系,找到适合不同数据集的参数.此外,文中的实验结果并不能显示出比对算法中哪类半监督聚类算法性能较好,这说明想要得到一种适用于大量数据集的聚类算法是有难度的.如何扩大算法的适应范围也是一个难题.为此,我们也将尝试将集成思想引入半监督聚类算法中,用集成算法来解决这一问题,这些想法是否能对聚类性能的进一步提升有所帮助还有待研究.我们下一步的工作是继续对参数选择问题和集成的半监督聚类算法进行探索性研究.

致谢 在此,我们特别感谢匿名审稿人提出的宝贵意见,这些意见对本文的修改有很大的帮助.

References:

[1] Demiriz A, Benneit KP, Embrechts MJ. Semi-Supervised clustering using genetic algorithm. In: Dagli CH, ed. Proc. of the Intelligent Engineering Systems Through Artificial Neural Networks (ANNIE'99). New York: ASME Press, 1999. 809−814.

[2] Bilenko M, Basu S, Mooney RJ. Integrating constraints and metric learning in semi-supervised clustering. In: Russ G, Dale S, eds. Proc. of the 21st Int'l Conf. on Machine Learning (ICML 2004). Banff: ACM Press, 2004. 81−88.

[3] Basu S, Bilenko M, Mooney RJ. A probabilistic framework for semi-supervised clustering. In: Won K, Ron K, Johannes G, William D, eds. Proc. of the 10th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining (KDD 2004). Seattle: ACM Press, 2004. 59−68.

[4] Wagstaff K, Cardie C. Clustering with instance-level constraints. In: Pat L, ed. Proc. of the 17th Int'l Conf. on Machine Learning (ICML 2000). Stanford: Morgan Kaufmann Publishers, 2000. 1103−1110.

[5] Wagstaff K, Cardie C, Rogers S, Schroedl S. Constrained *K*-means clustering with background knowledge. In: Carla EB, Andrea PD, eds. Proc. of the 18th Int'l Conf. on Machine Learning (ICML 2001). Williamstown: Morgan Kaufmann Publishers, 2001. 577−584.

[6] Basu S, Banerjee A, Mooney RJ. Semi-Supervised clustering by seeding. In: Claude S, Achim GH, eds. Proc. of 19th Int'l Conf. on Machine Learning (ICML 2002). Sydney: Morgan Kaufmann Publishers, 2002. 27−34.

[7] Kamvar SD, Klein D, Manning CD. Spectral learning. In: Georg G, Toby W, eds. Proc. of the 18th Int'l Joint Conf. on Artificial Intelligence (IJCAI 2003). Morgan Kaufmann Publishers, 2003. 561−566.

[8] Xu QJ, desJardins M, Wagstaf K. Constrained spectral clustering under a local proximity structure assumption. In: Ingrid R, Zdravko M, eds. Proc. of the 18th Int'l Florida Artificial Intelligence Research Society Conf. (FLAIRS 2005). AAAI Press, 2005. 866−867.

[9] Klein D, Kamvar SD, Manning CD. From instance-level constraints to space-level constraints: Making the most of prior knowledge in data clustering. In: Claude S, Achim GH, eds. Proc. of the 19th Int'l Conf. on Machine Learning (ICML 2002). Sydney: Morgan Kaufmann Publishers, 2002. 307−314.

[10] Wang L, Bo LF, Jiao LC. Density-Sensitive semi-supervised spectral clustering. Journal of Software, 2007,18(10):2412−2422 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/18/2412.htm

[11] Xing EP, Ng AY, Jordan MI, Russell S. Distance metric learning with application to clustering with side-information. In: Thrun S, Becker S, Obermayer K, eds. Advances in Neural Information Processing Systems (NIPS 2003). Cambridge: MIT Press, 2003. 505−512.

[12] Schultz M, Joachims T. Learning a distance metric from relative comparisons. In: Thrun S, Becker S, Obermayer K, eds. Advances in Neural Information Processing Systems (NIPS 2003). Cambridge: MIT Press, 2003. 40−47.

[13] Bar-Hillel A, Hertz T, Shental N, Weinshall D. Learning distance functions using equivalence relations. In: Tom F, Nina M, eds. Proc. of the 20th Int'l Conf. on Machine Learning (ICML 2003). Washington: AAAI Press, 2003. 11−18.

[14] Tang W, Xiong H, Zhong S, Wu J. Enhancing semi-supervised clustering: A feature projection perspective. In: Pavel B, Rich C, Xindong W, eds. Proc. of the 13th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining (KDD 2007). San Jose: ACM, 2007. 707−716.

[15] Erick CF, Zeidat N, Zhao ZH. Supervised clustering—Algorithms and benefits. In: Proc. of the 16th IEEE Int'l Conf. on Tools with Artificial Intelligence (ICTAI 2004). Boca Raton: IEEE Press, 2004. 774−776.

[16] Dettling M, Buhlmann P. Supervised clustering of genes. Genome Biology, 2002,3(12):research0069.1−0069.15.

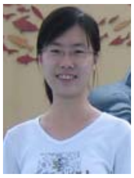
[17] Frey BJ, Dueck D. Clustering by passing messages between data points. Science, 2007,315(5814):972−976.

[18] Mézard M. Where are the exemplars? Science, 2007,315(5814):949−951.

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

附中文参考文献:

[10] 王玲,薄列峰,焦李成.密度敏感的半监督谱聚类.软件学报,2007,18(10):2412−2422. http://www.jos.org.cn/1000-9825/18/2412.htm



肖宇(1983－),女,河北保定人,博士生,CCF学生会员,主要研究领域为机器学习,数据挖掘.

于剑(1969－),男,博士,教授,博士生导师,CCF高级会员,主要研究领域为机器学习,数据挖掘.

第4届数字媒体及其在博物馆和文化遗产中的应用(DMAMH 2009)国际会议

征 文 通 知

第4届数字媒体及其在博物馆和文化遗产中的应用(DMAMH 2009)国际会议将于2009.7.25～2009.7.27在中国青岛举行。会议将由中国图像图形学会虚拟现实专委会主办，山东科技大学承办，中国国家自然科学基金委员会协办。本次会议的目的是为数字媒体、博物馆、多媒体领域的研究人员提供一个最新技术知识和经验的交流平台，并讨论今后相关领域未来研究的方向。届时将有该领域世界著名专家作报告，会议论文集将由IEEE出版(EI检索)。同时，本次大会将向Springer LNCS Transactions on Edutainment(EI检索)和International Journal of Virtual Reality(美国出版)等期刊推荐一些优秀论文。

一、会议主题(包括但不限于此)

数字博物馆	古文献数字化	图像/模型/视频水印	三维重建
虚拟博物馆导航	文物复原	图像分割	基于图像的绘制
遗产的建模和绘制	基于遥感技术的考古	多媒体数据库	实时图形渲染
虚拟遗产	地理信息系统	多媒体技术	动画技术
文物保护信息技术	虚拟现实/增强现实技术	图片/模型检索	交互技术和设备
文物计算机辅助评价	基于Web的展示	几何造型	媒体艺术

二、投稿须知

1．会议论文只接受英文稿件，必须是未曾发表过的研究结果，不要超过8页。来稿请按IEEE格式，使用A4或者8 1/2"×11"的稿纸，稿件内容格式为单栏，10号~12号字体。论文的内容包括论文题目、作者姓名以及联系方式（包括传真和电子邮件），另外还需提供一篇不超过200字的摘要。

2．所有论文都要经过至少两位专家的评审。主要从论文观点的准确性和原创性、研究成果的意义和影响、论文观点阐述的质量等方面评审论文。

3．注意：请将您要提交的论文稿件发至dmamh2009@sdust.edu.cn，或者dmamh2009@yahoo.com.cn

三、重要日期

征文截止日期：2009年4月10日	早期注册日期：2009年7月15日之前
稿件录用通知日期：2009年5月1日	现场注册日期：2009年7月25日~7月27日
修改稿截止日期：2009年5月10日	会议日期：2009年7月25日~7月27日