DOI: 10.11992/tis.201703013
网络出版地址: http://kns.cnki.net/kcms/detail/23.1538.TP.20170702.1547.026.html

结合稀疏表示与约束传递的半监督谱聚类算法

赵晓晓, 周治平
（江南大学 物联网技术应用教育部工程研究中心, 江苏 无锡 214122）

摘　要: 针对半监督谱聚类不能有效处理大规模数据, 没有考虑约束传递不能充分利用有限约束信息的问题, 提出一种结合稀疏表示和约束传递的半监督谱聚类算法。首先, 根据约束信息生成约束矩阵, 将其引入到谱聚类中; 然后, 将约束集合中的数据作为地标点构造稀疏表示矩阵, 近似获得图相似度矩阵, 从而改进约束谱聚类模型; 同时, 根据地标点的相似度矩阵生成连通区域, 在每个连通区域内动态调整近邻点, 利用约束传递进一步提高聚类准确率。实验表明, 所提算法和约束谱聚类相比, 在算法效率方面具有明显优势, 且准确率没有明显下降; 和快速谱聚类方法相比, 在聚类准确率上有所提升。

关键词: 数据挖掘; 聚类分析; 谱聚类; 半监督学习; 稀疏表示; 约束传递
中图分类号: TP18　　**文献标志码**: A　　**文章编号**: 1673−4785(2018)05−0855−09

中文引用格式: 赵晓晓, 周治平. 结合稀疏表示与约束传递的半监督谱聚类算法[J]. 智能系统学报, 2018, 13(5): 855–863.
英文引用格式: ZHAO Xiaoxiao, ZHOU Zhiping. A semi-supervised spectral clustering algorithm combined with sparse representation and constraint propagation[J]. CAAI transactions on intelligent systems, 2018, 13(5): 855–863.

A semi-supervised spectral clustering algorithm combined with sparse representation and constraint propagation

ZHAO Xiaoxiao, ZHOU Zhiping
(Engineering Research Center of Internet of Things Technology Applications Ministry of Education, Jiangnan University, Wuxi 214122, China)

Abstract: The semi-supervised spectral clustering algorithm does not deal with large-scale datasets effectively and does not fully utilize the constraint information because it does not consider the constraint propagation. To address these drawbacks, this paper proposes a semi-supervised spectral clustering algorithm that combines sparse representation and constraint propagation. The algorithm first generates the constraint matrix according to the constraint information, introduces it into the spectral clustering, and then constructs a sparse representation matrix by taking the data points in the constrained sets as the landmarks to approximate the graph similarity matrix, thereby revising the constrained spectral clustering model. Meanwhile, the connected region is generated according to the similarity matrix of the landmark data points, and the neighboring nodes are dynamically adjusted in each connected region. The clustering accuracy is further improved using the constraint propagation. Experimental results show that the proposed method is more efficient than constrained spectral clustering algorithms, and their accuracy levels are similar. Moreover, its clustering accuracy exceeds those of the fast spectral clustering algorithms.

Keywords: data mining; cluster analysis; spectral clustering; semi-supervised learning; sparse representation; constraint propagation

谱聚类作为聚类分析一种有效的方法, 建立在谱图划分的基础上, 可将数据集从原始空间转换到低维特征空间, 使原始数据变成线性可分[1],

收稿日期: 2017−03−10.　**网络出版日期**: 2017−07−02.
基金项目: 国家自然科学基金项目 (61373126).
通信作者: 赵晓晓. E-mail: 6151905019@vip.jiangnan.edu.cn.

目前已广泛应用于图像分割、人脸识别等领域[2-3]。另一方面，半监督学习是机器学习领域的研究热点，已被用于解决实际问题[4-6]，在聚类分析中引入一些监督信息来指导聚类过程，能够提高聚类准确率。

半监督聚类算法的约束信息包括“必连”和“勿连”约束集合，引入这些约束信息可指导聚类过程。目前，针对半监督谱聚类算法已有大量的研究，Kamvar 等[7]根据约束关系调整数据之间的相似度，将调整后的相似度矩阵用于改进谱聚类，但是不能充分利用初始有限的约束关系；蒋伟进等[8]提出一种纠错式主动学习成对约束算法，将挖掘到的监督信息用于调整数据点之间的距离矩阵，但约束集合对聚类结果影响较大；丁世飞等[9]通过优化高斯核参数和引入成对约束信息来调整相似度矩阵，但过多的约束信息会对聚类准确率造成负面的影响；Cucuringu 等[10]将“必连”和“勿连”约束矩阵均视为图拉普拉斯矩阵，把约束聚类转化为广义特征值问题，针对 2 类划分效率和准确率显著提高。王翔等[11]提出一种柔性约束谱聚类框架，引入大量硬约束和软约束信息建立新的约束优化问题，有效提高聚类准确率，但是不具备约束关系传递。基于约束信息的半监督聚类算法，通常能获取数据的约束关系是非常有限的，通过约束传递来获得大量可靠的成对约束信息，可显著提高半监督聚类的性能。余志文等[12]提出一种增强型半监督聚类集成框架 (incremental semi-supervised clustering ensemble, ISSCE)，采用卢志武等[13]提出的约束传递方法，即基于 k 近邻图的标签传播方法，将少量标签样本的监督信息传递给未标签样本，但如果相似度错误反映数据点之间的相似性，会造成约束关系的错误传递。传统的谱聚类及上述大部分半监督谱聚类算法均需要存储数据的相似度矩阵且对拉普拉斯矩阵进行特征分解，空间和时间的计算复杂度比较高，在处理大规模数据集时计算代价难以承受。为了提升谱聚类的扩展性，蔡登等[14]提出基于地标点表示的谱聚类算法，通过数据点与地标点之间的相似度矩阵乘积来近似得到整体数据点的相似度矩阵，然后利用近似性质实现快速特征分解。

本文将采用基于地标点近似表示的方法，通过稀疏表示构造的相似度矩阵对柔性约束谱聚类算法模型[11]进行改进，并且根据约束集合内地标点的关系，利用 Tarjan 算法[15]自动检测连通区域，在每个连通区域内部动态调整近邻点，传递约束关系，从而更新稀疏表示矩阵提高聚类准确率。在 5 个大规模标准数据集上进行实验的结果表明，本文算法对这些大规模数据具有较好适应性，且在有效降低算法复杂度的同时，保证了约束谱聚类算法结果的准确率。

1　算法基本原理

1.1　约束 NCut 算法

对于数据集 $\boldsymbol{X}=\{\boldsymbol{x}_i\}_{i=1}^n, \boldsymbol{x}_i\in\mathbf{R}^d$，数据之间的相似度矩阵为 $\boldsymbol{W}$，度矩阵 $\boldsymbol{D}$ 为对角阵，其内元素表示为 $D_{ii}=\sum_j W_{ij}$，规范化拉普拉斯矩阵表示为 $\boldsymbol{L}=\boldsymbol{I}-\boldsymbol{D}^{-1/2}\boldsymbol{W}\boldsymbol{D}^{-1/2}$，$\boldsymbol{I}$ 表示单位矩阵。

标准 NCut 的目标函数为

$$\arg\min_{\boldsymbol{v}\in\mathbb{R}^n}\boldsymbol{v}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{v},\ \text{s.t.}\boldsymbol{v}^{\mathrm{T}}\boldsymbol{v}=\boldsymbol{1},\boldsymbol{v}\perp\boldsymbol{1} \tag{1}$$

式中 $\boldsymbol{v}$ 表示松弛化的类指示向量。

假设已知“必连”约束集合 M 与“勿连”约束集合 C，文献[11]根据约束信息生成约束矩阵 $\boldsymbol{Q}$，可表示为

$$Q_{ij}=\begin{cases}1, & (\boldsymbol{x}_i,\boldsymbol{x}_j)\in M\\ -1, & (\boldsymbol{x}_i,\boldsymbol{x}_j)\in C\\ 0, & \text{无约束信息}\end{cases} \tag{2}$$

通过式 (3) 来衡量 $\boldsymbol{Q}$ 中约束关系与 $\boldsymbol{v}$ 的一致程度：

$$\boldsymbol{v}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{v}=\sum_{i=1}^n\sum_{j=1}^n\boldsymbol{v}_i\boldsymbol{v}_jQ_{ij} \tag{3}$$

定义软约束 $\boldsymbol{v}\in\mathbf{R}^n$，$\boldsymbol{Q}\in\mathbf{R}^{n\times n}$，如果数据 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$ 是同一类，那么 $\boldsymbol{Q}$ 为正；否则为负。$\boldsymbol{Q}$ 的绝对值表示约束权重，$\boldsymbol{v}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{v}$ 的值越大，表明 $\boldsymbol{v}$ 与 $\boldsymbol{Q}$ 中的约束信息就越一致。基于上述软约束，不一定要严格满足每个指定的约束条件，忽略部分约束信息可降低计算开销，通过用户指定阈值来确定约束的下界值，即 $\boldsymbol{v}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{v}\geqslant\alpha$。

约束 NCut 的目标函数为

$$\arg\min_{\boldsymbol{v}\in^n}\boldsymbol{v}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{v},\ \text{s.t.}\boldsymbol{v}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{v}\geqslant\alpha,\boldsymbol{v}^{\mathrm{T}}\boldsymbol{v}=\boldsymbol{1},\boldsymbol{v}\perp\boldsymbol{1} \tag{4}$$

上述问题可转化为求解矩阵的特征向量[11]，但是空间和时间复杂度分别为 $O(n^2)$ 和 $O(n^3)$，对于大规模数据集，计算复杂度过高；而且不具备约束传递功能，不能充分利用有限的约束信息。

1.2　依据稀疏表示的相似度矩阵

蔡登等[14]提出基于图稀疏表示的谱聚类算法，通过地标点的线性组合来实现所有数据点 $\boldsymbol{X}$ 的近似表示 $\boldsymbol{X}\approx\boldsymbol{Y}\boldsymbol{Z}$。对于给定的任一数据点 $\boldsymbol{x}_i$，其近似数据点 $\hat{\boldsymbol{x}}_i$ 可以表示为

$$\hat{\boldsymbol{x}}_i=\sum_{j=1}^p Z_{ji}\boldsymbol{y}_j \tag{5}$$

式中：$\boldsymbol{y}_j$ 是 $\boldsymbol{Y}\in\mathbf{R}^{d\times p}$ 的列向量，Z_{ji} 是 $\boldsymbol{Z}\in\mathbf{R}^{p\times n}$ 的第 j 行第 i 列元素。

如果$\boldsymbol{x}_i$越接近$\boldsymbol{y}_j$, Z_{ji}会越大, 如果$\boldsymbol{y}_j$不在数据点$\boldsymbol{x}_i$的 r 近邻内, 可设Z_{ji}为 0, 所以可生成一个稀疏表示矩阵 $\boldsymbol{Z}$。$\boldsymbol{Y}_{\langle i\rangle}\in\mathbf{R}^{d\times r}$表示 $\boldsymbol{Y}$ 的子矩阵, 由$\boldsymbol{x}_i$的 r 近邻组成, Z_{ji}计算见式 (6):

$$Z_{ji}=\frac{K(\boldsymbol{x}_i,\boldsymbol{y}_j)}{\sum_{j'\in\langle i\rangle}K(\boldsymbol{x}_i,\boldsymbol{y}_{j'})},\quad i\in 1,2,\cdots,n,j\in\langle i\rangle \tag{6}$$

式中: $K(\cdot)$是核函数, 较常用的是高斯核函数$K(\boldsymbol{x}_i,\boldsymbol{y}_j)=\mathrm{e}^{\frac{-\|x_i-y_j\|}{2\sigma^2}}$; σ表示带宽。

在获得矩阵 $\boldsymbol{Z}$ 后, 可以构造两种形式的图, 即$\boldsymbol{G}=\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z}$和$\boldsymbol{S}=\boldsymbol{Z}\boldsymbol{Z}^{\mathrm{T}}$, 计算$\hat{\boldsymbol{Z}}=\boldsymbol{D}^{-1/2}\boldsymbol{Z}$, $\boldsymbol{D}$ 为对角矩阵, 其内元素$D_{ii}=\sum_j Z_{ij}$, 所以规范化图的相似度矩阵可表示为$\hat{\boldsymbol{G}}=\hat{\boldsymbol{Z}}^{\mathrm{T}}\hat{\boldsymbol{Z}}\in\mathbf{R}^{n\times n}$和$\hat{\boldsymbol{S}}=\hat{\boldsymbol{Z}}\hat{\boldsymbol{Z}}^{\mathrm{T}}\in\mathbf{R}^{p\times p}$, 计算$\hat{\boldsymbol{G}}$的时间复杂度为$O(pn^2)$, 计算$\hat{\boldsymbol{S}}$的时间复杂度为$O(np^2)(p\ll n)$。

2 所提算法

2.1 约束关系传递

将在“必连”和“勿连”约束集合中的数据点视为地标点, 其个数为 p。根据约束关系生成一个$p\times p$的相似度矩阵, 利用 Tarjan 算法检测强连通区域, 会生成多个连通区域; 对每个连通区域, 动态调整其内数据的近邻点, 进行约束关系的传递, 对基于地标点近似表示方法中的稀疏表示矩阵$\hat{\boldsymbol{Z}}$进行更新。

根据约束信息所生成的相似度矩阵, 利用 Tarjan 算法检测生成$|H|$个强连通区域, 可表示为$T\equiv\{T_1\cup T_2\cup\cdots\cup T_a\}$, 其中$a\in\{1,2,\cdots,|H|\}$, $|T|=p$, T_a表示第 a 个连通区域, 包含$T_{|a|}$个数据点, T'_a表示T_a的补集, $T'_a\equiv\{p'_a\in T|p'_a\notin T_a\}$, 同一个连通区域内的数据之间相似度最大, 在不同连通区域的数据之间的相似度最小, 设置为

$$\hat{\boldsymbol{Z}}(p_i,p_j)=\begin{cases}1, & \{p_i\in T_a|p_j\in T_a\}\\ 0, & \{p_i\in T_a|p_j\in T'_a\}\end{cases} \tag{7}$$

每个数据点均有 r 个近邻点, $\boldsymbol{L}_{T_a}\in\mathbf{R}^{\mathbf{r}\times|\mathbf{T_a}|}$表示在$T_a$内每个数据点与其近邻点之间的稀疏表示矩阵$\hat{\boldsymbol{Z}}$集合, 依次计算每个近邻点$i\in\{1,2,\cdots,r\}$的出现次数$\mathrm{freq}_i$, 并且按照升序进行排列, 生成一个 m 维的向量$\mathbf{degreeVec}=(\mathrm{freq}_{1'},\mathrm{freq}_{2'},\cdots,\mathrm{freq}_{m'})$, m 是$\boldsymbol{L}_{T_a}$近邻点出现次数不同的个数, 出现次数较多的近邻点是$p_a\in T_a$最近的邻居点, 依据线性插值策略, 生成一个 m 维的 **lookupVec** 向量:

$$\mathbf{lookupVec}_{i'}=\begin{cases}\min+\mathrm{freq}_{i'}\times\dfrac{\max-\min}{m-1}, & m\neq 1\\ \max, & m=1\end{cases} \tag{8}$$

式中: min 和 max 分别表示$\boldsymbol{L}_{T_a}$中最大和最小相似度, $i'\in\{1,2,\cdots,m\}$。

$\boldsymbol{L}_{p_i}$和$\boldsymbol{L}_{p_j}$分别表示在同一个连通区域内的数据p_i和p_j的 r 近邻集合, $p_i,p_j\in T_a$, $i,j\in\{1,2,\cdots,|T_a|\}$。因为矩阵 $\boldsymbol{Z}$ 具有高度稀疏性, 在同一个连通区域内的两个数据点p_i和p_j可能并没有共同近邻点。所以可考虑对于T_a内的任一数据点p_i, 将其他数据点$p_j\in T_a$的近邻点也作为p_i的近邻点:

$$\boldsymbol{L}_{p_i}=\left\{\boldsymbol{L}_{p_1}\cup\boldsymbol{L}_{p_2}\cup\cdots\cup\boldsymbol{L}_{p_{|a|}}\right\},\quad i\in 1,2,\cdots,|T_a| \tag{9}$$

根据式 (8)、(9) 进行约束关系的传递, 对矩阵$\hat{\boldsymbol{Z}}$进行更新:

$$\hat{\boldsymbol{Z}}(i,p_j)=\mathbf{lookupVec}_{i'} \tag{10}$$

式中: $i\in\boldsymbol{L}_{p_j}$, $p_j\in T_a$, $\mathrm{freq}_i=\mathrm{freq}_{i'}$, i 是p_j的近邻点, 它出现的频率freq_i等于按照升序排列后的频率$\mathrm{freq}_{i'}$, 存储在向量**degreeVec**中。

2.2 依据稀疏表示的可扩展半监督 NCut 算法

根据 1.2 节所提出的基于稀疏表示建立的相似度矩阵可知, $\hat{\boldsymbol{G}}=\hat{\boldsymbol{Z}}^{\mathrm{T}}\hat{\boldsymbol{Z}}$是数据 $\boldsymbol{X}$ 的规范化相似度矩阵, 经过 2.1 节的约束关系传递, 已经实现对矩阵$\hat{\boldsymbol{Z}}$的更新, 相似度矩阵$\hat{\boldsymbol{G}}$也随着更新, 且包含更多的约束信息, 能够有效地提高聚类结果。

依据稀疏表示的约束 NCut 目标函数可表示为

$$\arg\min_{\boldsymbol{v}\in\mathbb{R}^n}\boldsymbol{v}^{\mathrm{T}}\bar{\boldsymbol{L}}\boldsymbol{v},\ \mathrm{s.t.}\boldsymbol{v}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{v}\geqslant\alpha,\boldsymbol{v}^{\mathrm{T}}\boldsymbol{v}=\mathbf{1},\boldsymbol{v}\perp\mathbf{1} \tag{11}$$

式中$\bar{\boldsymbol{L}}=\boldsymbol{I}-\hat{\boldsymbol{G}}\in\mathbf{R}^{n\times n}$是规范化的图拉普拉斯矩阵。

基于文献[11]提出的方法, 上述问题的求解可以松弛为一般的特征值求解问题:

$$\boldsymbol{L}\boldsymbol{v}=\lambda(\boldsymbol{Q}-\beta\boldsymbol{I})\boldsymbol{v} \tag{12}$$

式中β是α的下界。

求解式 (12) 特征向量的时间复杂度为$O(n^3)$, 所以在处理大规模数据集时计算负担过大。为了使该方法具有可扩展性, 能较好地适应于大规模数据集, 根据 1.2 可知, 基于稀疏表示的方法还可以构造另外一种规范化图, 其相似度矩阵表示为$\hat{\boldsymbol{S}}=\hat{\boldsymbol{Z}}\hat{\boldsymbol{Z}}^{\mathrm{T}}\in\mathbf{R}^{p\times p}$, 若能将$\hat{\boldsymbol{S}}$引入到上述约束 NCut 的目标函数中, 可以大大降低求解特征向量的时间复杂度。因此可将$\boldsymbol{v}\in\mathbf{R}^n$表示为$\hat{\boldsymbol{Z}}^{\mathrm{T}}\boldsymbol{u}$, 其中$\boldsymbol{u}\in\mathbf{R}^p$。将$\boldsymbol{v}=\hat{\boldsymbol{Z}}^{\mathrm{T}}\boldsymbol{u}$代入到式 (11)。改进后的可扩展约束 NCut 目标函数可表示为

$$\arg\min_{\boldsymbol{u}\in\mathbb{R}^p}\boldsymbol{u}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{u},\ \mathrm{s.t.}\boldsymbol{u}^{\mathrm{T}}\hat{\boldsymbol{Q}}\boldsymbol{u}\geqslant\alpha,\boldsymbol{u}^{\mathrm{T}}\hat{\boldsymbol{S}}\boldsymbol{u}=\mathbf{1},\mathbf{1}^{\mathrm{T}}\hat{\boldsymbol{S}}\boldsymbol{u}=0 \tag{13}$$

式中: $\boldsymbol{A}=\hat{\boldsymbol{S}}-\hat{\boldsymbol{S}}\hat{\boldsymbol{S}},\hat{\boldsymbol{Q}}=\hat{\boldsymbol{Z}}\boldsymbol{Q}\hat{\boldsymbol{Z}}^{\mathrm{T}}$。

该模型等价于式 (11) 所表示的约束 NCut 目标函数, 但是有两个改进: 一是规范化的拉普拉斯矩阵$\boldsymbol{L}\in\mathbf{R}^{n\times n}$变为矩阵$\boldsymbol{A}\in\mathbf{R}^{p\times p}$; 二是约束矩阵$\boldsymbol{Q}\in\mathbf{R}^{n\times n}$变为$\hat{\boldsymbol{Q}}\in\mathbf{R}^{p\times p}$, 因为$p\ll n$, 有效降低了计算复杂度。同样和文献[11]所提出的约束 Ncut 模型相比, 一方面, 通过地标点所构造的稀疏表示矩阵来近似获得相似度矩阵, 避免对整个数据进行

特征分解，大大降低算法的复杂度，能够很好地适应于大规模数据集；另一方面，在连通区域内部进行约束关系传递，更新矩阵$\hat{\boldsymbol{Z}}$，改进模型即式(13)中$\hat{\boldsymbol{S}}$和$\boldsymbol{A}$也随着更新，可以充分利用初始有限的约束信息，提高聚类的准确率。

为了求解式(13)，引入拉格朗日乘子，可扩展约束NCut问题转化为

$$\Lambda(\boldsymbol{u},\lambda,\mu)=\boldsymbol{u}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{u}-\lambda\left(\boldsymbol{u}^{\mathrm{T}}\hat{\boldsymbol{Q}}\boldsymbol{u}-\alpha\right)-\mu\left(\boldsymbol{u}^{\mathrm{T}}\hat{\boldsymbol{S}}\boldsymbol{u}-\boldsymbol{1}\right) \tag{14}$$

需要满足KKT条件，即

$$\begin{gathered}\boldsymbol{A}\boldsymbol{u}-\lambda\hat{\boldsymbol{Q}}\boldsymbol{u}-\mu\hat{\boldsymbol{S}}\boldsymbol{u}=0 \\ \boldsymbol{u}^{\mathrm{T}}\hat{\boldsymbol{Q}}\boldsymbol{u}\geqslant\alpha \\ \boldsymbol{u}^{\mathrm{T}}\hat{\boldsymbol{S}}\boldsymbol{u}=\boldsymbol{1} \\ \boldsymbol{1}^{\mathrm{T}}\hat{\boldsymbol{S}}\boldsymbol{u}=0 \\ \lambda\geqslant 0 \\ \lambda\left(\boldsymbol{u}^{\mathrm{T}}\hat{\boldsymbol{Q}}\boldsymbol{u}-\alpha\right)=0\end{gathered} \tag{15}$$

当$\lambda=0$时，式(15)变为$\boldsymbol{A}\boldsymbol{u}-\mu\hat{\boldsymbol{S}}\boldsymbol{u}=0$，意味着并没有考虑相关约束信息，所以为了充分使用约束信息，只考虑$\lambda>0$的情况，即$\boldsymbol{u}^{\mathrm{T}}\hat{\boldsymbol{Q}}\boldsymbol{u}=\alpha$。

假设$\beta=-\dfrac{\mu}{\lambda}$，式(15)变为

$$\boldsymbol{A}\boldsymbol{u}=\lambda\left(\hat{\boldsymbol{Q}}-\beta\hat{\boldsymbol{S}}\right)\boldsymbol{u} \tag{16}$$

通过求解式(16)的特征值和特征向量，计算复杂度远远降低。

还需要考虑如何设置参数β和α，由于矩阵$\boldsymbol{A}$是半正定的，可得

$$\lambda\boldsymbol{u}^{\mathrm{T}}\left(\hat{\boldsymbol{Q}}-\beta\hat{\boldsymbol{S}}\right)\boldsymbol{u}=\lambda(\alpha-\beta)\geqslant 0$$

即参数β是α的下界值，α作为约束的下界值，所以只需要指定参数β值即可，参数β可调意味着算法对噪声等额外信息的处理更加灵活。为了保证式(16)有i个有意义的特征向量，需要满足$\beta<\gamma_i$，其中γ_i表示$\hat{\boldsymbol{Q}}x=\gamma\hat{\boldsymbol{S}}x$的第$i$个最大的特征值。

算法 结合稀疏表示与约束传递的半监督谱聚类算法

输入 数据集$\boldsymbol{X}$，约束集合M和C，聚类个数k，参数β；

输出 类label。

1) 将在约束集合内的p个点选为地标点，并作为矩阵$\boldsymbol{U}$的列向量；

2) 根据式(6)构造$\boldsymbol{Z}$，并计算$\hat{\boldsymbol{Z}}=\boldsymbol{D}^{-1/2}\boldsymbol{Z}$；

3) 使用Tarjan算法，根据地标点之间的约束关系计算连通区域CC；

4) for 每个连通区域CC do：

计算连通区域的邻接子矩阵$\boldsymbol{L}_{T_a}$；

根据近邻点出现频率的次数构建向量**lookupVec**；

根据式(9)将约束关系传递给该区域内剩余数据点；

根据**lookupVec**更新$\hat{\boldsymbol{Z}}$；

end for

5) 依据约束关系生成约束矩阵$\boldsymbol{Q}$，计算$\hat{\boldsymbol{S}}=\hat{\boldsymbol{Z}}\hat{\boldsymbol{Z}}^{\mathrm{T}}$，$\hat{\boldsymbol{Q}}=\hat{\boldsymbol{Z}}\boldsymbol{Q}\hat{\boldsymbol{Z}}^{\mathrm{T}}$；

6) 求解$\hat{\boldsymbol{Q}}x=\gamma\hat{\boldsymbol{S}}x$的最大特征值$\gamma_{\max}$；

7) if $\beta\geqslant\gamma_{\max}$，类label的向量$\boldsymbol{V}=\varnothing$；

8) else 求解式(16)的特征向量$\boldsymbol{u}$；

9) 找出所有正特征值所对应的特征向量$\{\boldsymbol{u}_i^+\}$，$\{\boldsymbol{u}_i^+\}$中每个元素乘$\sqrt{\dfrac{1}{\boldsymbol{u}_i\hat{\boldsymbol{S}}\boldsymbol{u}_i}}$；

10) 去掉$\{\boldsymbol{u}_i^+\}$中与$\boldsymbol{1}^{\mathrm{T}}\hat{\boldsymbol{S}}$不正交的特征向量；

11) 计算$\boldsymbol{u}_i^{\mathrm{T}}\boldsymbol{A}\boldsymbol{u}_i$，寻找$m$个特征向量使$\boldsymbol{u}_i^{\mathrm{T}}\boldsymbol{A}\boldsymbol{u}_i$最小，$m=\min\left\{k-1,\left|\{\boldsymbol{u}_i^+\}\right|\right\}$，并把这些特征向量作为$\boldsymbol{V}$的列向量；

12) 计算$\boldsymbol{V}^{(r)}=\hat{\boldsymbol{Z}}^{\mathrm{T}}\boldsymbol{V}(\boldsymbol{I}-\boldsymbol{V}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{V})$，并对其进行$k$-means聚类。

2.3 算法复杂度分析

基于地标点近似表示的谱聚类算法需要构造稀疏矩阵$\boldsymbol{Z}$，该步骤的时间复杂度为$O(pn)$，对矩阵$\boldsymbol{Z}$进行奇异值分解获得相似度矩阵的特征向量，该步骤的时间复杂度为$O(p^3+p^2n)$；所提算法也要构造矩阵$\boldsymbol{Z}$，生成$|H|$个连通区域进行近邻点约束关系传递的时间复杂度为$O(r|H|)$，计算矩阵$\hat{\boldsymbol{S}}$的时间复杂度为$O(p^2n)$，求解式(16)的特征值计算复杂度为$O(p^3)$，远远小于文献[11]约束谱聚类算法特征分解所需要的时间复杂度$O(n^3),p\ll n$，计算$\hat{\boldsymbol{Z}}\boldsymbol{Q}\hat{\boldsymbol{Z}}^{\mathrm{T}}$的时间复杂度为$O(kp^2+kpn)$。

3 实验与分析

3.1 实验环境

为了验证本文算法的性能，选取规模较大的数据集进行实验，依次为物体图像数据集COIL100，人脸数据集CMU-PIE，手写数字数据集USPS、MNIST和UCI标准库中的森林植被类数据集CoverType，数据集的特性如表1所示。仿真实验基于MATLAB2014b平台，计算机的硬件配置为Intel i7-4770 CPU 3.40 GHz、16 GB RAM。

表 1 实验数据集的特性

Table 1 Experimental datasets features

数据集	数据集个数	类	维数
COIL100	7 200	100	1 024
USPS	9 298	10	256
CMU-PIE	11 554	68	1 024
MNIST	70 000	10	784
CoverType	581 012	7	54

本文算法和文献[11]的约束谱聚类方法 CSP, 文献[14]的基于地标点采样的快速谱聚类算法 LSC-R 和 LSC-K, 文献[15]所提出的针对大规模数据集的半监督谱聚类算法 SC-PC 进行对比, LSC-R 是通过随机采样来获取 p 个地标点, LSC-K 是利用 k-means 算法来获得 p 个地标点。须指出, 由于 CSP 算法在 CMU-PIE、MNIST 和 CoverType 数据集上计算负担过大, 并没有进行相关实验的比较。

鉴于比较的公平性, 具体的参数设置如下: 其中文献[14-15]和本文算法的地标点个数 p 均设置为 1 000, 所有算法中 k-means 部分的迭代次数均设置为 500, 所有稀疏表示矩阵构造过程中的近邻点个数 r 设置为 5。约束的下界即参数$\beta=\beta_0\gamma_{k-1}$, 其中 $\beta_0=0.5+0.4\times\frac{p}{n}$, γ_{k-1}为$\hat{\boldsymbol{Q}}x=\gamma\hat{\boldsymbol{S}}x$的第$k-1$个最大的特征值。

通过数据集中每个样本预定义的类标签来实现对聚类结果的评价, 采用聚类准确率 ACC 和归一化互信息 NMI 两种度量指标[14]对聚类结果进行评估和比较分析。两个评价指标的取值范围均是在 0~1 之间, 值越大表示聚类效果越好。

3.2　实验分析

各种算法在上述数据集的实验结果如表 2 所示。

根据表 2 可以看出, 约束谱聚类算法 CSP 在 COIL100 和 USPS 两个数据集上的 ACC 和 NMI 均取得最佳结果, 因为其考虑引入约束矩阵建立约束优化问题, 提高聚类准确率, 但是在上述两个数据集上的运行时间太长, 耗时将近 5~7 h, 对于更大的数据集 CMU-PIE、MNIST 和 CoverType 数据集, 计算负担过大, 没能进行验证。本文算法在 COIL100 和 USPS 数据集的 ACC 分别为 0.666 0 和 0.843 5, NMI 指标分别为 0.797 5 和 0.773 1, 和 CSP 相比稍有降低; 但从运行时间上, 本文算法的运行时间仅在秒级别, 耗时分别为 3.51 s 和 1.88 s, 远远少于 CSP 的运行时间, 而且在包含 581 012 个样本的大规模 CoverType 数据集上运行时间只需要 747.36 s, 所以本文算法基于稀疏表示矩阵来建立相似度矩阵, 降低了矩阵分解的复杂度, 提高了半监督聚类算法的可扩展性。

另一方面, 本文算法和 LSC-K、LSC-R、SC-PC 三种均为提升谱聚类的效率的算法, 即快速谱聚类算法相比, 在准确率 ACC 和归一化互信息 NMI 两个指标上有所提高, 且在 CMU-PIE 和 CoverType 数据集上两个指标具有明显的提升; 和 SC-PC 算法相比, 本文算法在 CMU-PIE 数据集上 ACC、NMI 分别提升了 79.12% 和 38.58%, 在 CoverType 数据集上 ACC 和 NMI 分别提升了 15.02% 和 39.94%。因为在处理高维数据时, 采用稀疏表示能够过滤一些异常点和噪声, 同时指定约束下界, 可去除一些噪声等约束关系的负面影响。因此, 引入约束信息来改变谱聚类算法的目标函数, 并通过在连通区域内进行约束关系传递, 能够有效提高聚类的准确率。

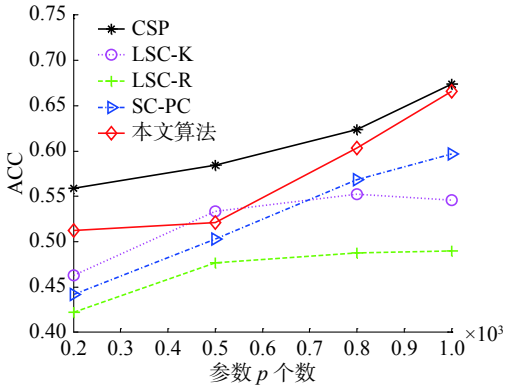
表 2　各数据集实验结果

Table 2　Experimental results of different datasets

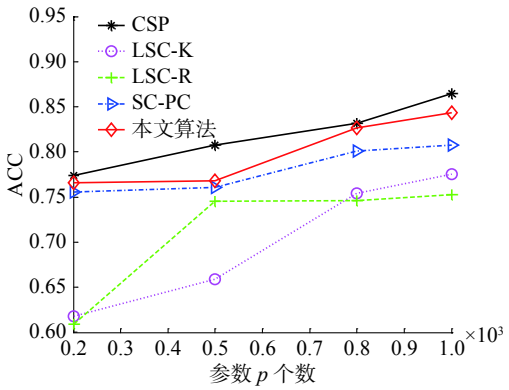
数据集	算法	ACC	NMI	时间/s
COIL100	CSP	0.673 9	0.823 7	19 656.02
	LSC-K	0.545 6	0.763 2	4.13
	LSC-R	0.489 6	0.731 5	2.33
	SC-PC	0.596 8	0.771 8	2.59
	本文	0.666 0	0.797 5	3.51
USPS	CSP	0.864 4	0.804 9	26 932.75
	LSC-K	0.775 3	0.791 5	1.19
	LSC-R	0.752 4	0.754 1	0.62
	SC-PC	0.807 4	0.769 7	0.75
	本文	0.843 5	0.773 1	1.88
CMU-PIE	CSP	—	—	—
	LSC-K	0.095 6	0.212 6	5.39
	LSC-R	0.093 6	0.214 8	4.48
	SC-PC	0.172 4	0.262 3	4.84
	本文	0.308 8	0.363 5	6.45
MNIST	CSP	—	—	—
	LSC-K	0.727 0	0.722 2	12.58
	LSC-R	0.589 0	0.591 1	8.80
	SC-PC	0.734 4	0.697 8	11.08
	本文	0.783 9	0.672 8	24.98
CoverType	CSP	—	—	—
	LSC-K	0.255 2	0.067 3	355.11
	LSC-R	0.229 0	0.068 1	78.18
	SC-PC	0.256 3	0.072 1	149.71
	本文	0.294 8	0.100 9	747.36

为了分析初始约束信息对聚类结果的影响, 添加了地标点个数 p(对应半监督聚类中能获取的初始约束信息) 对聚类指标影响的对比实验。同样, CSP 算法只在 USPS 和 COIL100 两个数据集

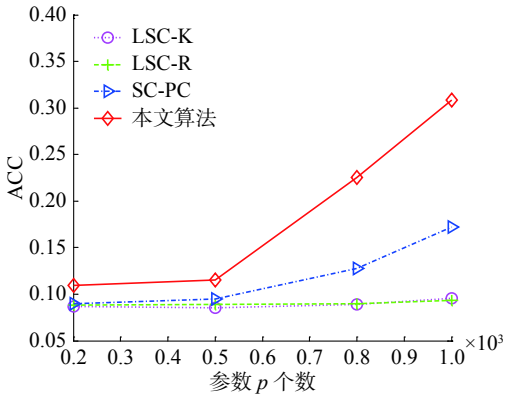
上实现。p 依次取 200、500、800 和 1 000。不同算法 ACC 和 NMI 的对比实验结果分别如图 1、2 所示。



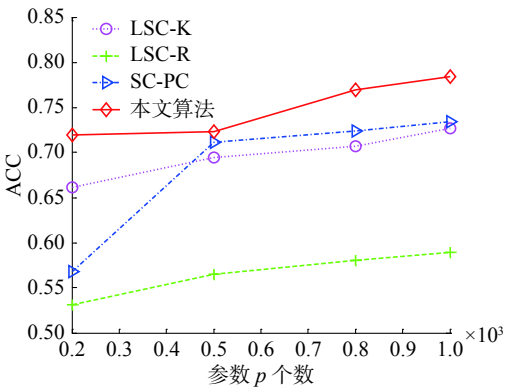
(a) COIL100 数据集



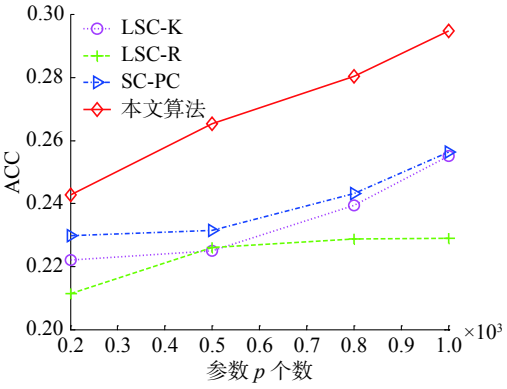
(b) USPS 数据集



(c) CMU-PIE 数据集



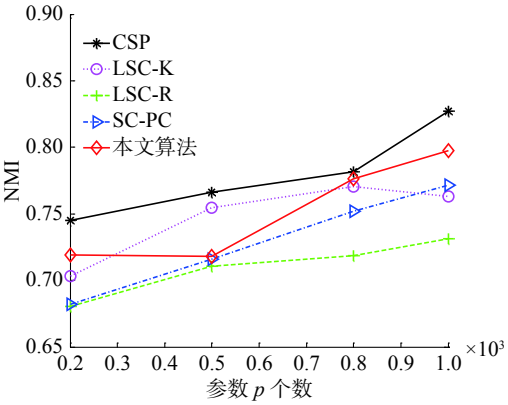
(d) MNIST 数据集



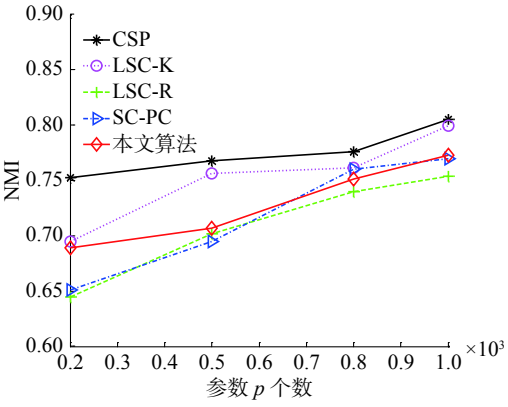
(e) CoverType 数据集

图 1　不同算法的 ACC 比较

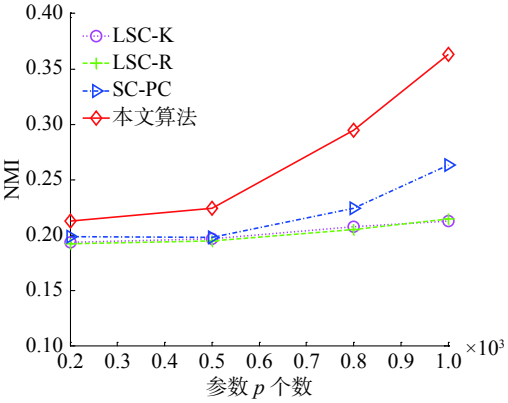
Fig. 1　Comparison of the ACCs of different algorithms



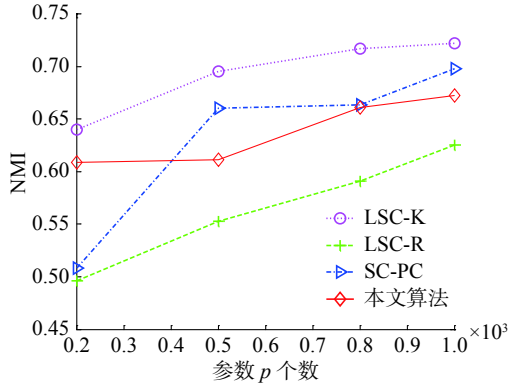
(a) COIL100 数据集



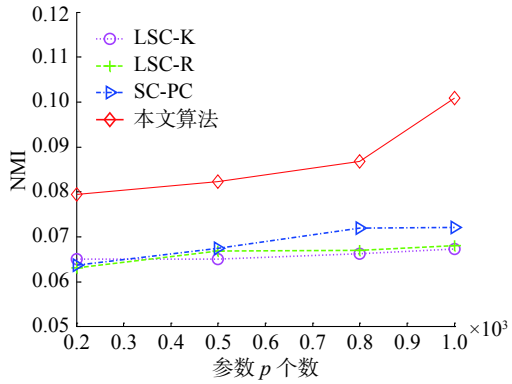
(b) USPS 数据集



(c) CMU-PIE 数据集



(d) MNIST 数据集



(e) CoverType 数据集

图 2　不同算法的 NMI 比较

Fig. 2　Comparison of the NMIs of different algorithms

由图 1 可知，本文算法在 5 个数据集上的 ACC 比 LSC-K、LSC-R 和 SC-PC 三种算法都要高，随着参数 p 个数增加，每种算法的 ACC 也随着增加，说明 ACC 受 p 的选取影响很大。由图 2 可知，在 NMI 指标上，本文算法和快速谱聚类算法相比，在 CMU-PIE 和 CoverType 两个数据集上具有明显的优越性，同样随着 p 个数增多，每种算法的 NMI 都随着提高。总的来说，在 p 的不同取值情况下，本文算法能取得最好的聚类效果。

4　结束语

随着规模庞大、结构复杂数据的不断出现，对其聚类往往耗费大量的时间，同时多数半监督谱聚类算法存在没有充分利用初始约束信息的问题。本文通过稀疏表示改进约束谱聚类算法的目标函数，根据初始约束信息生成连通区域，在每个连通区域内动态调整近邻点进行约束关系传递。实验结果表明，本文算法在提高聚类准确率的同时能有效降低聚类复杂度，能够较好地适用于大规模数据集。但是本文算法的聚类准确率受采样地标点的个数和选取方法影响比较大，接下来可以针对该问题开展进一步的研究工作。

参考文献：

[1] VON Luxburg U. A tutorial on spectral clustering[J]. Statistics and computing, 2007, 17(4): 395–416.

[2] SHI Jianbo, MALIK J. Normalized cuts and image segmentation[J]. IEEE transactions on pattern analysis and machine intelligence, 2000, 22(8): 888–905.

[3] HU Han, FENG Jianjiang, YU Chuan, et al. Multi-class constrained normalized cut with hard, soft, unary and pairwise priors and its applications to object segmentation[J]. IEEE transactions on image processing, 2013, 22(11): 4328–4340.

[4] 刘建伟, 刘媛, 罗雄麟. 半监督学习方法[J]. 计算机学报, 2015, 38(8): 1592–1617.
LIU Jianwei, LIU Yuan, LUO Xionglin. semi-supervised learning methods[J]. Chinese journal of computers, 2015, 38(8): 1592–1617.

[5] ALUSH A, FRIEDMAN A, Goldberger J. Pairwise clustering based on the mutual-information criterion[J]. Neurocomputing, 2016, 182: 284–293.

[6] FORESTIER G, WEMMERT C. Semi-supervised learning using multiple clusterings with limited labeled data[J]. Information sciences, 2016, 361-362: 48–65.

[7] KAMVAR S D, KLEIN D, MANNING C D. Spectral learning[C]//Proceedings of the 18th International Joint Conference on Artificial Intelligence. Acapulco, Mexico, 2003: 561–566.

[8] 蒋伟进, 许宇晖, 王欣. 基于谱图和成对约束的主动半监督聚类算法[J]. 控制与决策, 2013, 28(6): 904–908.
JIANG Weijin, XU Yuhui, WANG Xin. Active semi-supervised clustering algorithm based-on pair-wise constraints[J]. Control and decision, 2013, 28(6): 904–908.

[9] DING Shifei, JIA Hongjie, ZHANG Liwen, et al. Research of semi-supervised spectral clustering algorithm based on pairwise constraints[J]. Neural computing and applications, 2014, 24(1): 211–219.

[10] CUCURINGU M, KOUTIS I, CHAWLA S, et al. Simple and scalable constrained clustering: a generalized spectral method[C]//Proceedings of the 19th International Conference on Artificial Intelligence and Statistics. Cadiz, Spain, 2016: 445–454.

[11] WANG Xiang, QIAN Buyue, DAVIDSON I. On constrained spectral clustering and its applications[J]. Data mining and knowledge discovery, 2014, 28(1): 1–30.

[12] YU Zhiwen, LUO Peinan, YOU J, et al. Incremental semi-supervised clustering ensemble for high dimension-

al data clustering[J]. IEEE transactions on knowledge and data engineering, 2016, 28(3): 701–714.

[13] LU Zhiwu, PENG Yuxin. Exhaustive and efficient constraint propagation: a graph-based learning approach and its applications[J]. International journal of computer vision, 2013, 103(3): 306–325.

[14] CAI Deng, CHEN Xinlei. Large scale spectral clustering via landmark-based sparse representation[J]. IEEE transactions on cybernetics, 2015, 45(8): 1669–1680.

[15] SEMERTZIDIS T, RAFAILIDIS D, STRINTZIS M G, et al. Large-scale spectral clustering based on pairwise constraints[J]. Information processing and management, 2015, 51(5): 616–624.

作者简介：

赵晓晓，女，1993 年生，硕士研究生，主要研究方向为数据挖掘。

周治平，男，1962 年生，教授，博士，主要研究方向为智能检测、自动化装置、网络安全。发表学术论文 80 余篇。

机电一体化、控制和机器人技术国际会议（ICMCR 2019）
2019 2nd International Conference on Mechatronics, Control and Robotics (ICMCR 2019)

2019 2nd International Conference on Mechatronics, Control and Robotics will be held in Fukuoka Institute of Technology, Fukuoka, during February 23–25, 2019. It aims to provide a forum for researchers, practitioners, and professinals from the industry, academia and government who are working in the field of mechatronics, control and robotics to discourse on research and development, professional practice in related fields.

ICMCR 2019 is also the Annual Meeting of JOACE editorial board, so it also serves to bring authors and editors of JOACE together to communicate face to face and discuss their latest research results and future development of JOACE.

Best Presentation Award: Selection of the best paper will be made at the conference based on both the technical content and presentation. The winner will be chosen by the Session Chair in consultaton with the Conference Chairs and the Best Presentation Certificate will be awarded.

可信网络交易系统的理论创新与实践推进
——《网络交易风险控制理论》书评

王怀清

香港城市大学信息系统学系, 中国　香港

随着电子商务的迅猛发展, 网络交易遭受恶意攻击、网络钓鱼以及交易欺诈等愈加严重。网络交易平台的关键技术、监管能力及手段明显不足。蒋昌俊教授领衔的科研团队对可信网络交易系统的理论和实践进行了多年深入的研究, 取得了一系列可喜的研究成果, 并收于《网络交易风险控制理论》一书。该书对网络交易系统的现状进行了深入总结, 对学术界和工业界常用的可信保障技术进行了全面介绍, 并提出了具有“行为”特色的认证机制, 以及一系列的基于模型的风险防控方法。该书内容丰富、资料翔实、逻辑严密, 是基于信息学科研究网络交易风险防控的一本好书。

近年来, 随着网络技术的发展, 以及“互联网+”相关政策的支持, 网络交易作为新的商业模式发展异常迅速。据中国互联网络信息中心统计, 截至 2017 年 12 月, 我国网民规模达 7.72 亿, 其中, 网络购物用户达 5.53 亿, 相较 2016 年增长了 14.3%, 占网民总体的 69.1%; 使用网上支付的用户规模也达 5.31 亿。网络交易的持续高速发展也带了诸多的安全问题, 比如恶意攻击、木马劫持等对网络交易造成了巨大的伤害, 其中, 2017 年遇到网络钓鱼以及各类交易欺诈的用户占比较 2016 年有所升高, 快递服务和电商网站相关维权搜索量占比最大。据 360 猎网平台分析, 金融理财诈骗是举报数量最多的类型, 虚假购物紧随其后。针对上述问题, 蒋昌俊教授及其团队经过多年的研究, 在相关项目的支持下, 取得了有价值的成果, 并汇总成《网络交易风险控制理论》一书, 该书研究内容系统全面, 研究特色凸显, 是目前笔者读到的基于信息学科研究网络交易风险控制的第一本专著。

《网络交易风险控制理论》研究内容系统全面, 由蒋昌俊教授和于汪洋博士合著, 于 2018 年 3 月出版, 该书从信息技术角度, 介绍了网络交易风险防控的理论和相关技术, 首次将行为认证引入网络交易的可信保障, 开展了行为认证技术在网络交易系统中的应用研究, 形成了网络交易支付系统风险防控关键技术的整套理论与方法, 并研制了大规模网络交易风险防控系统平台。

全书共分为 8 章。第 1 章介绍了目前国内外网络交易的现状, 对学术界、业界常用的可信保障技术进行了全面的总结, 并提出了相关问题, 指出传统的安全技术和身份认证方法已不足以应对开放网络环境下新的风险挑战, 研究基于行为的认证理论有望取得创新性的成果。第 2 章则介绍了该书中用到的领域内基本知识, 为后面章节的展开进行铺垫, 其中, 形式化模型是重要的一环, 因为要对并发的网络交易系统行为进行准确地刻画, Petri 网、自动机等形式化模型必不可少。对于网络交易这一并发软件系统, 其行为是依托于业务流程的, 所以对于网络交易流程的刻画至关重要, 因而该书第 3 章着重介绍了业务系统的相关知识, 并突出了作者团队提出的基于 Petri 网的两种形式化模型, 这些成果均已发表于国内外知名期刊, 具有一定的影响力。第 4 章、第 5 章则开始涉及基于行为的认证理论, 针对软件系统的行为认证和用户的行为认证, 作者提出了大量创新性的理论和方法, 融合了服务器端和客户端的可信认证, 同时也涉及测试技术、系统评估和身份认证等现有的风险防控技术。基于上述理论, 作者研究团队开发了大规模网络交易风险防控系统平台, 监控技术是核心。第 6 章介绍了监控平台的框架、核心技术和异常处理机制等。

随着国内互联网金融经济的发展, 互联网征信受到越来越多的重视。征信也是风险防控的重要一环, 对用户信用的掌握可以有效减少网络交易的风险。因此, 第 7 章主要介绍了征信系统的基本原理和架构。最后, 作者用了 3 个案例研究来结束该书。总的来说, 该书内容丰富, 覆盖全面, 结构清晰, 逻辑通顺, 将网络交易风险控制的现状、理论、方法进行了科学的整理和阐述。

该书研究特色凸显, 一大特色是提出了网络交易风险防控的行为认证技术, 建立了交易系统行为和用户行为的规范与模式, 通过采集和分析用户在系统中的行为足迹, 建立了基于模型的行为认证机制, 有效提高

了网络交易风险防控的实时性，降低了误判率，克服了交易欺诈的高辨识和强实时的难题。其中，行为认证的核心技术在于形式化模型的构建。在当前人工智能和大数据的背景下，使用形式化方法进行建模和分析必不可少，基于严格的数学定义和分析可以最大程度地发现问题并解决问题，不仅能发现已知的缺陷，还能发现未知的缺陷，为风险防控的高效实施提供模型基础。蒋昌俊教授及其团队在形式化领域耕耘多年，有着深厚的理论积淀，尤其在 Petri 网的理论和应用方面，成就斐然，培养了一大批杰出人才。将适于刻画并发系统的 Petri 网应用于网络交易系统的风险防控，水到渠成。同时，该书也不拘泥于 Petri 网这一种形式化模型，根据不同的应用场景，行为认证技术也涉及了自动机、马尔科夫模型等广泛使用的理论和方法，为网络交易的风险防控奠定了坚实的理论基础。

除了理论上的创新，该书的研究还进行了工业实践方面的探索，并卓有成效。据我所知，蒋昌俊教授领衔的团队在进行上述理论研究的过程中，得到了国家自然科学基金委和科技部多个项目的支持，并与国内领先的网络交易平台 (支付宝、快钱、中国工商银行等单位) 建立了良好的合作关系。对于支付宝等网络交易企业来说，在进行海量交易的同时，如何高效准确地对每笔交易进行风险判别，是亟待解决的问题，而传统的以数字证书为核心的安全技术已不足以应对。该书的研究恰好契合了业界的需要，作者所在团队的科研人员与支付宝等知名企业进行了长期的合作研究，部分人员常驻支付宝，将理论研究的成果进行实践应用，与企业共同对风控系统进行升级和优化，攻克了大规模、高并发、强实时交易的若干关键问题，并搭建了大规模分布式的网络交易风险防控平台。这是产学研结合的不错尝试，是在企业业务场景和技术基础上，借力高校的理论创新优势，向前迈进的一步，对交易风险的精准判定和瞬时识别具有重要的意义。

网络交易已成为国民经济的重要组成部分，其中的安全可信问题越发凸显，如何在技术层面有效控制风险是一个关键的科学问题。该著作的研究紧紧契合这一科学问题，将网络交易风险控制的现状、理论、方法进行了科学的整理和阐述，既有现有的成熟技术，又有独创的理论方法，对互联网金融风险控制具有很强的启发、借鉴和指导作用，同时对信息技术领域的相关研究人员、学者也具有参考价值，对计算机软件理论与软件安全专业的学生也不失为一本开拓视野的参考书。

笔者非常愿意将此书推荐给对网络交易风险防控感兴趣的学者及工业界。同时，感谢作者在这一领域所进行的研究，期盼作者能够出更多的优秀成果。