doi:10.3969/j.issn.1671-637X.2016.05.007

基于聚类集成的半监督多/高光谱图像分类方法

吕俊伟[1], 樊利恒[1,2], 邓江生[2], 石晓航[1]
(1.海军航空工程学院控制工程系,山东 烟台 264001; 2.海军航空装备计量监修中心,上海 200436)

摘 要:提出一种基于聚类集成的半监督多/高光谱图像分类方法。谱聚类是近年来出现的基于图论的、以相似性为基础的一类性能优越的聚类算法,能对任意形状分布的数据进行聚类,但对参数的变化比较敏感。聚类集成技术可有效提高单聚类算法的精度和稳定性,并具有良好的鲁棒性和泛化能力。算法利用聚类集成算法的优点并利用谱聚类的思想开发聚类集成算法的共识函数,将谱聚类作为聚类成员来构造聚类集成系统,使用高斯RBF核映射下的多维数据的光谱角制图计算权值矩阵W,并用Nyström方法来降低算法的运算复杂度,实现了多/高光谱遥感数据的半监督分类。最后通过实验验证了该算法无论对多光谱还是高光谱都有较好的分类效果。

关键词:图像分类;半监督分类;多光谱图像;高光谱图像;谱聚类;聚类集成
中图分类号:TP391 **文献标志码**:A **文章编号**:1671-637X(2016)05-0030-07

Semi-supervised Classification of Multi/Hyperspectral Images Based on Cluster Ensemble

LYU Jun-wei[1], FAN Li-heng[1,2], DENG Jiang-sheng[2], SHI Xiao-hang[1]
(1. Department of Control Engineering, Naval Aeronautical and Astronautical University, Yantai 264001, China;
2. Naval Aeronautical Equipment Measure, Supervise and Maintenance Center, Shanghai 200436, China)

Abstract: A semi-supervised, cluster-ensemble based method for the classification of multi/hyperspectral images is presented. Spectral clustering is a graph theory based clustering algorithm taking similarity as the basis, and has become increasingly popular in recent years. It can deal with arbitrary distribution of dataset but with a drawback for being sensitive to the scaling parameters. Cluster ensemble techniques are effective in improving both the robustness and the stability of the single clustering algorithm. Cluster ensemble also has a character of good robustness and generalization ability. The processing method in this paper utilizes the merits of cluster ensemble and develops a consensus function based spectral clustering algorthm. The clustering components are generated by spectral clustering. The affinity matrix is generated by computing the SAM between different datapoints. The Nyström method is used to to speed up the classification process. Thus semi-supervised classification to multi/hyperspectral remote sensed data is completed. Experiments show that the method presented here has an excellent classificaation result for both multispectral and hyperspectral remote sensed dataset.

Key words: image classification; semi-supervised classification; multispectral image; hyperspectral image; spectral clustering; cluster ensemble

0 引言

在高光谱图像分类中,监督分类策略理论上能够获得所能得到的最高分类精度[1-2],但在学习训练阶段需要大量的标签样本来构造精确的分类器。随着成像技术的进步,以较小代价获得海量数据变得十分容易,但获取标签样本的代价仍然较高。非监督分类也称聚类分析,其分类结果只是对不同类别达到了区分,并不能确定类别的属性[3-4]。针对此类问题,半监督分类方法应运而生并得到了广泛的研究[5-9],它是介于监督分类和无监督分类之间的分类技术,克服了监督分类方法

收稿日期:2015-04-10 **修回日期**:2016-03-15
基金项目:国家自然科学基金(61032001,60801049);国家"八六三"计划创新基金(2010AAJ140)
作者简介:吕俊伟(1960—),男,山东牟平人,博士,教授,研究方向为遥感图像处理、机器视觉、目标识别与跟踪。

不容易获得足够数目的训练样本和非监督分类不能确定类别属性的缺点。半监督学习可以利用未标签样本所含对分类有帮助的信息来修正学习过程。当在实际应用中只有少量的标签数据和大量的未标签数据时，利用半监督学习策略进行分类一般能提供一个比较满意的分类结果，半监督学习还具有一定的泛化能力，从而使训练得到的分类器具有更好的性能。

谱聚类是近年来出现的一类性能优越的聚类算法[10-13]，具有识别非高斯分布的能力，能对任意形状的数据进行聚类。文献[14]将集成学习引入到聚类分析中，提出了聚类集成算法，与单一的聚类算法相比，聚类集成具有更好的鲁棒性、适用性、稳定性、并行性和扩展性。

本文提出一种基于聚类集成的半监督多/高光谱图像分类方法，并用 Nyström 方法来降低算法的运算复杂度，最后对实际多/高光谱遥感图像进行分类实验，实验结果显示本文算法对多光谱遥感图像和高光谱遥感图像都取得了较好的分类效果。

1　谱聚类和聚类集成

1.1　谱聚类

由于谱聚类算法具有识别非高斯分布的能力，适合于许多实际问题，它能对任意形状的数据进行聚类，其思想源于谱图划分，将数据聚类问题看成是一个无向图的多路划分问题。数据点看作是无向图的顶点，边代表数据集合中数据点之间的相似性。令 $G(V,E)$ 是一个无向图，它的顶点为数据点集合 V，由一组边界 E 相连，连接顶点 i 和 j 的边界相应的权重为 W_{ij}，则可由全部数据点构建权值矩阵 $\boldsymbol{W}$。谱聚类的主要工具是图的拉普拉斯矩阵，规范拉普拉斯矩阵的形式为

$$\boldsymbol{L}=\boldsymbol{I}-\boldsymbol{D}^{-1/2}\boldsymbol{W}\boldsymbol{D}^{-1/2}=\boldsymbol{I}-\boldsymbol{S} \tag{1}$$

式中：$\boldsymbol{D}$ 为对角阵，$D_{ii}=\sum_j W_{ij}$；$\boldsymbol{S}$ 常被称为相似度矩阵。

目前，谱聚类算法已被成功应用于并行计算、超大规模集成电路设计、图像分割、数据挖掘等领域，但是仍存在以下两个主要缺点：1）对尺度参数比较敏感，使得相似度矩阵 $\boldsymbol{S}$ 的构造比较困难；2）需要求解矩阵的特征值分解问题，对于大规模应用，其计算量和存储量太大，造成矩阵的特征值和特征向量难以计算。例如，对于一个 145×145 的单色图像，存储图像像素点之间的相似性需要 21025×21025 大小的矩阵。

1.2　聚类集成

在聚类集成中，先要产生对数据集 X 的 M 个聚类成员，然后对这 M 个成员的聚类结果根据某个准则进行合并，因此，聚类集成研究主要包括如何产生具有适当差异度的聚类成员，以及如何设计共识函数这两个方面[15-17]。

通常可以将聚类集成问题表述如下：令 $X=\{x_1,x_2,\cdots,x_n\}$ 表示由 n 个数据点组成的集合，$\Pi=\{\pi_1,\pi_2,\cdots,\pi_M\}$ 表示 X 进行 M 次聚类得到的聚类结果，其中，$\pi_i=\{C_1^i,C_2^i,\cdots,C_{ki}^i\}$ 为第 i 次聚类得到的簇集合，k_i 表示第 i 个簇集合中簇的个数。对数据点 $x\in X$，$C(x)$ 表示其簇标签，在第 i 个簇集合中，如果 $x\in C_j^i$，则有 $C(x)=j$。聚类集成就是对集合 Π 进行合并，得到数据集 X 最终的聚类结果。

1.3　半监督分类

高光谱遥感图像分类包括监督分类和非监督分类。其中：监督分类方法（如人工神经网络）在处理非常多的波段时效率较差，而基于核的方法和支持向量机在处理高光谱图像的分类时有成功的应用，而且能够以一种鲁棒的方式处理样本噪声，但是，监督分类面临的主要困难是学习过程严重依赖于训练数据的质量，而只有在同步或同样的条件下得到同样类别的图像中才能满足，更坏也是经常会出现的情况是只能得到数目受限或者完全得不到训练样本；非监督分类方法在高光谱图像分类中已经展示出良好的分类结果，非监督分类作用在整幅图像上，因而对标签样本的数目不敏感，但聚类和类别的关系不能确定。

半监督分类是介于监督分类和无监督分类之间的一种分类方法，半监督学习方法可以同时利用标签样本和未标签样本数据来构造分类器，因而半监督分类自然而然地能够提供更好的分类结果。当在实际应用中只有少量的标签样本和大量的未标签样本数据时，利用半监督学习策略进行分类一般能提供一个更满意的分类结果，而且可以提高分类器的泛化能力，从而进一步提高了分类器的性能。一般来说，目前文献中常见的高光谱图像半监督分类方法有 3 种：1）生成模型法，需要估计条件概率密度函数，已经广泛应用于遥感图像的分类；2）低密度分离算法，该算法同时最大化标签样本和未标签样本之间的边界，最近开始应用于高光谱图像的分类；3）基于图的方法，每个样本都向其“邻居”传播它的类别信息直到整个数据集达到一种稳定的状态。

鉴于基于图的半监督学习方法具有很好的数学解释性、与核方法之间的关系和良好的学习性能，而谱聚类算法的思想源于谱图划分，本文提出了基于谱聚类的聚类集成半监督高光谱分类方法。

2　基于聚类集成的半监督分类算法

2.1　聚类成员的构造

将单个谱聚类算法作为聚类成员。谱聚类算法对尺度参数非常敏感，不同的参数可能会得到完全不同的

聚类结果,对于谱聚类来说这是一个明显的缺陷,而多样性已被证明是提高聚类集成性能的关键因素,因此谱聚类的这个缺陷对于聚类集成来说则是一个优点。

设给定的由高光谱遥感像素组成的高维数据集为 $X=\{x_1,\cdots,x_l,x_{l+1},\cdots,x_n\}\subset R^N$,$N$ 为光谱波段数,$x_1,\cdots,x_l$ 为标签样本,$x_{l+1},\cdots,x_n$ 为未标签样本,$L=\{1,\cdots,c\}$ 为标签集,对于 $x_i(i\leqslant l)$,它们的标签 $y_i\in L$。令 $\mathscr{F}$ 为一个 $n\times c$ 的非负矩阵,$\boldsymbol{F}=[\boldsymbol{F}_1^{\mathrm{T}},\cdots,\boldsymbol{F}_n^{\mathrm{T}}]$,$\boldsymbol{F}\in\mathscr{F}$ 表示数据集 X 上的一个分类,每个样本 $x_i\in X$ 的标签 $y_i=\mathrm{argmax}_{j\leqslant c}F_{ij}$。定义一个 $n\times c$ 的矩阵 $\boldsymbol{Y}\in\mathscr{F}$,如果 x_i 被标记为 $y_i=j$,则 $Y_{ij}=1$,否则 $Y_{ij}=0$。即

$$\begin{cases}Y_{ij}=1 & \text{if} \quad y_i=j\\ Y_{ij}=0 & \text{otherwise}\end{cases}。\tag{2}$$

聚类 $\pi=\{C_1,C_2,\cdots,C_c\}$ 中簇 C_i 的样本个数为 $\mathrm{sum}(y_{ij})$,簇 C_i 中的数据点集合为 $X_i=\{x_i|x_i\times y_{ij}=x_i\}$。

算法总结如下所述。

1) 使用高斯 RBF 核映射下的光谱角制图计算权值矩阵 $\boldsymbol{W}$,即

$$W_{ij}\equiv\boldsymbol{W}(x_i,x_j)=\exp\frac{-S_{\mathrm{SAM}}(x_i,x_j)}{(2\sigma^2)}\quad\forall i\neq j\tag{3}$$

$$S_{\mathrm{SAM}}(x_i,x_j)=\arccos\frac{x_i\cdot x_j}{\|x_i\|\ \|x_j\|}\tag{4}$$

光谱角制图(Spectral Angle Mapper,SAM)是对地物光谱波形相似性的一种度量[18]。为了避免自相似,定义 $W_{ij}=0$。

2) 构造相似度矩阵 $\boldsymbol{S}$,即

$$\boldsymbol{S}=\boldsymbol{D}^{-1/2}\boldsymbol{W}\boldsymbol{D}^{-1/2}\tag{5}$$

式中,$\boldsymbol{D}$ 为对角阵,$D_{ii}=\sum_j W_{ij}$。这一步对应于特征空间的正规化。考虑通过映射后样本的点积形式形成的半正定核矩阵,$\boldsymbol{W}=\langle\phi(x_i),\phi(x_j)\rangle$,则正规化形式为

$$\hat{W}(x_i,x_j)=\left[\frac{\phi(x_i)}{\|\phi(x_i)\|},\frac{\phi(x_j)}{\|\phi(x_j)\|}\right]=\frac{\boldsymbol{W}(x_i,x_j)}{\sqrt{\boldsymbol{W}(x_i,x_i)\boldsymbol{W}(x_j,x_j)}}\tag{6}$$

则可以得到一个聚类结果,即一个聚类成员。

3) 构造一系列聚类成员,即

$$\boldsymbol{F}(m+1)=\alpha\boldsymbol{S}\boldsymbol{F}(m)+(1-\alpha)\boldsymbol{Y}\quad m=1,\cdots,M-1\tag{7}$$

$$\boldsymbol{F}(M)|_{M\to\infty}=(1-\alpha)(\boldsymbol{I}-\alpha\boldsymbol{S})^{-1}\boldsymbol{Y}\tag{8}$$

式中,参数 α 的取值范围为[0,1],α 取不同的值,可以得到一系列的聚类成员。

2.2 聚类集成的实现

在得到聚类成员后,需要设计共识函数以得到最终的聚类结果。总体来说,主要有以下 3 类共识函数:1) 基于特征的方法;2) 基于图的方法;3) 基于数据点间相似度的方法。方法 3) 建立数据点之间的互关联矩阵并用基于相似度聚类的方法来得到聚类结果,容易产生具有链式结构的簇。文献[19]提出了一种新的基于连接的相似度矩阵构造方法,即连接三元组算法(Connected Triple Algorithm)。

连接簇 C_i 和簇 C_j 的边的权重为 W_{Cij},由两个簇共同包含的数据点个数得到

$$W_{Cij}=|X_i\cap X_j|/|X_i\cup X_j|\tag{9}$$

式中,X_i 为属于簇 C_i 的数据点的集合,$X_i=\{x_i|x_i\times y_{il}=x_i\}$。临界点为 C_k 的两个簇 C_i 和 C_j 之间的连接三元组的个数,用 WCT_{ij}^k 表示为

$$WCT_{ij}^k=\min(w_{ik},w_{jk})\tag{10}$$

簇 C_i 和 C_j 之间的相似度为

$$Sim^{WCT}(i,j)=\frac{\sum_{k=1}^{q}WCT_{ij}^k}{WCT_{\max}}\tag{11}$$

式中,$1\leqslant q<\infty$。

对任意的簇集合 $\pi_m\in\Pi$,$m=1,\cdots,M$,数据点 x_i,$x_j\in X$ 之间的相似度为

$$\boldsymbol{S}_m(x_i,x_j)=\begin{cases}1 & \text{if} \quad C(x_i)=C(x_j)\\ Sim^{WCT}(C(x_i),C(x_j)) & \text{otherwise}\end{cases}\tag{12}$$

式中,$\boldsymbol{S}$ 即为数据点之间的相似度矩阵。

2.3 Nyström 方法

一个减少谱聚类和聚类集成中相似度矩阵计算复杂度及存储量大的问题的途径是只保留正规化相似度矩阵最大的 p 个特征值,但只有在 $p\ll n$ 时才能较大幅度地减少计算复量,而 $p\ll n$ 可能会造成精度的降低。

本文利用 Nyström 方法对相似度矩阵进行分解,可以有效解决谱聚类和聚类集成中相似度矩阵计算量和存储量大的问题。

对于谱聚类,相似度矩阵 $\boldsymbol{S}$ 的大小为 $n\times n$,n 是标签样本和未标签样本的总数。通过从原始的相似度矩阵 $\boldsymbol{S}$ 中随机选取 m 行/列形成一个新的近似矩阵 $\widetilde{\boldsymbol{S}}$。

$$\widetilde{\boldsymbol{S}}_{n,n}=\boldsymbol{S}_{n,m}\boldsymbol{S}_{m,m}^{-1}\boldsymbol{S}_{m,n}\quad m\leqslant n\tag{13}$$

式中,$\boldsymbol{S}_{n,m}$ 表示矩阵 $\boldsymbol{S}$ 的一个 $n\times m$ 块。

$$\boldsymbol{S}=\boldsymbol{D}^{-1/2}\boldsymbol{W}\boldsymbol{D}^{-1/2}=\boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^{\mathrm{T}}\tag{14}$$

式中:$\boldsymbol{V}$ 为对应于矩阵 $\boldsymbol{S}$ 特征向量的归一化矩阵;$\boldsymbol{\Lambda}$ 为对角线元素为 $\boldsymbol{S}$ 特征值的对角阵。则

$$\widetilde{\boldsymbol{S}}_{n,n}=\widetilde{\boldsymbol{V}}\widetilde{\boldsymbol{\Lambda}}\widetilde{\boldsymbol{V}}^{\mathrm{T}}\tag{15}$$

则对于谱聚类,算法复杂度从 $O(n^3N)$ 变为 $O(mn^2N)$,其中,N 为光谱波段数。对于聚类集成,算法复杂度从从 $O(M^3)$ 变为 $O(mM^2)$。

如果以张成大小为 $p\times p$ 的小矩阵 $\widetilde{\boldsymbol{S}}=\widetilde{\boldsymbol{V}}\widetilde{\boldsymbol{\Lambda}}\widetilde{\boldsymbol{V}}^{\mathrm{T}}$ 来近似代替 $\boldsymbol{S}$,并代入算式 $\boldsymbol{F}(m+1)=\alpha\boldsymbol{S}\boldsymbol{F}(m)+(1-\alpha)\cdot\boldsymbol{Y}$,则

$$\boldsymbol{F}(m+1)=\alpha\widetilde{\boldsymbol{V}}\widetilde{\boldsymbol{\Lambda}}\widetilde{\boldsymbol{V}}^{\mathrm{T}}\boldsymbol{F}(m)+(1-\alpha)\boldsymbol{Y}\tag{16}$$

式中,$m=1,\cdots,M-1$。

当 $M\to\infty$ 时，有

$$\boldsymbol{F}(m+1)=\alpha\,\widetilde{\boldsymbol{V}}\widetilde{\boldsymbol{\Lambda}}\,\widetilde{\boldsymbol{V}}^{\mathrm{T}}\boldsymbol{F}(m)+(1-\alpha)\boldsymbol{Y} \quad (17)$$

$$\boldsymbol{F}(M)\,|_{M\to\infty}=(1-\alpha)(\boldsymbol{I}-\alpha\,\widetilde{\boldsymbol{V}}\widetilde{\boldsymbol{\Lambda}}\,\widetilde{\boldsymbol{V}}^{\mathrm{T}})^{-1}\boldsymbol{Y} \quad (18)$$

根据线性代数中的 Woodbury 方程式

$$(C+AB)^{-1}=C^{-1}-C^{-1}A(\boldsymbol{I}+BC^{-1}A)^{-1}BC^{-1} \quad (19)$$

可以得到

$$F(M)\,|_{M\to\infty}=(1-\alpha)(Y-\widetilde{V}(\widetilde{\Lambda}\widetilde{V}^{\mathrm{T}}\widetilde{V}-\alpha^{-1}I)^{-1}\widetilde{\Lambda}\widetilde{V}^{\mathrm{T}}Y) \quad (20)$$

此时，对于谱聚类和聚类集成的计算复杂度分别为 $O(p^2nN)$ 和 $O(p^2M)$，是样本数 n 或聚类数 M 的线性函数，大大减小了计算复杂度。

3　实验与分析

3.1　实验数据

1）实验数据 1。

该数据是多光谱数据 TipJul1. lan，该图像拍摄于 1986 年 7 月 16 日，是位于美国印第安纳州 Tippecanoe 郡的一幅场景。图像共有 7 个波段，将该数据的第 4 个波段作为红色，第 3 个波段作为绿色，第 2 个波段作为蓝色的合成颜色，显示如图 1 所示，包含的类别有 Corn，Soybean，Wheat，Alfalfa/Oats，Pasture 和 Hay/Grassland 等，如图 2 所示，其中，Corn，Soybean 和 Alfalfa/Oats 的面积较大，作为初始阶段感兴趣的类别。该多光谱数据图和地面观测数据图源于 http://dynamo. ecn. purdue. edu/~biehl/MultiSpec，该图中 Corn，Soybean 和 Alfalfa/Oats 的样本数目分别为 9437，7589 和 2284。

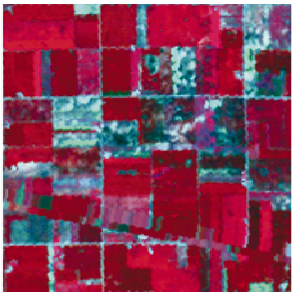


图 1　原始遥感图像（波段 4，3，2）

Fig. 1　Original RSI（band 4，3，2）

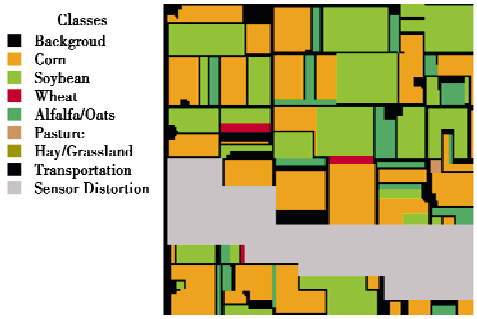


图 2　参考分类图像

Fig. 2　The reference classification image

2）实验数据 2。

该数据是美国 AVIRIS（Airborne Visible InfraRed Imaging Spectrometer）高光谱遥感数据 92AV3C. tif，采集自美国印第安纳州西北部的一块印度松树测试地，该高光谱遥感数据是公开的基准高光谱图像数据。将该数据的第 50 个波段作为红色，第 27 个波段作为绿色，第 17 个波段作为蓝色的合成颜色，显示如图 3 所示，其地面真实数据如图 4 所示。从图中可以看出，该数据包括 16 个地物类别。

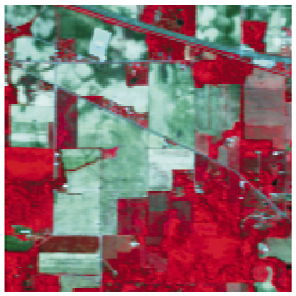


图 3　AVIRIS 的合成颜色显示图（波段 50，27，17）

Fig. 3　The synthetical display image of 92AV3C. tif (band 50，27，17)

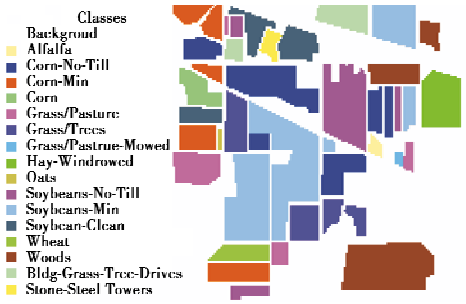


图 4　92AV3C. tif 的地面观测数据图

Fig. 4　A generalized reconnaissance map of the dataset

高光谱数据和地面观测数据图来源于 http://dynamo. ecn. purdue. edu/~biehl/MultiSpec，数据的地物类别情况如表 1 所示。由于序号为 1，4，7，9，13，15 和 16 的类别可用样本数较少，这里只考虑了序号为 2，3，5，6，8，10，11，12 和 14 的类别为实验对象，并分别用 C_2，C_3，C_5，C_6，C_8，C_{10}，C_{11}，C_{12} 和 C_{14} 代替类别名称。

表 1　实验 2 所用高光谱遥感数据的地物类别及样本数目

Table 1　The classes and number of samples in the experiments

类别名称	地物类别	样本点数目	类别名称	地物类别	样本点数目
C_1	Alfalfa	54	C_9	Oats	20
C_2	Corn-No-Till	1434	C_{10}	Soybeans-No-Till	968
C_3	Corn-Min	834	C_{11}	Soybean-Min	2468
C_4	Corn	234	C_{12}	Soybean-Clean	614
C_5	Grass/Pasture	497	C_{13}	Wheat	212
C_6	Grass/Trees	747	C_{14}	Woods	1294
C_7	Grass/Pasture-Mowed	26	C_{15}	Bldg-Grass-Tree-Drives	380
C_8	Hay-Windrowed	489	C_{16}	Stone-Steel Towers	95

3.2 实验与分析

本文实验软件环境为 Matlab 2008a，Windows XP，硬件环境为 Intel Celeron E3300 CPU 2.51 GHz/1.99 GB RAM。

1）参数的选取。

在使用谱聚类方法构造聚类成员时，用高斯 RBF 核计算由光谱角制图表征的两个样本点相似度，高斯 RBF 核为 $W_{ij} \equiv W(x_i, x_j) = \exp \dfrac{-S_{\mathrm{SAM}}(x_i, x_j)}{(2\sigma^2)}$，$\sigma$ 的取值为 $\{10^{-3}, \cdots, 10^{3}\}$。

在使用 $\boldsymbol{F}(m+1) = \alpha \boldsymbol{S}\boldsymbol{F}(m) + (1-\alpha)\boldsymbol{Y}$ 构造聚类成员时，α 的取值为 $\{0.01, \cdots, 0.99\}$。

2）样本的选取。

在已知类别的样本中随机选取一部分作为标签样本，隐藏剩余样本标签信息，将其作为未标签样本。标签样本的数目对最终的分类结果应该有影响，可以期待，随着标签样本数目的增加，分类精度和 kappa 系数应该有不同程度的增加，这里各类标签样本数目分别为 {10，20，50，100}。选取的标签样本的质量对最终的分类结果也有影响，对每种标签样本进行 10 次实验，并取最好的一次作为结果显示。

3）分类精度评价。

分类精度评价是分类技术中一个必不可少的环节。进行精度评价，一方面可以有效地对分类算法进行评估从而改善分类算法，另一方面也是对分类成果的最终评价。精确分析分类器的分类精度是一件非常复杂和困难而又难以使众多不同分析者信服的事情，更多信息可以参考文献[5]。

由于选择的样本实际上都是已知类别的样本，只是大部分在分类中临时隐藏了其类别标签，因此评价分类精度是非常方便的。

使用最广泛的表示分类精度的方式是构建一个 $k \times k$ 的混淆矩阵 $\boldsymbol{A}$，k 表示类别数，a_{ij} 表示分类结果中第 i 类与参考类型数据第 j 类所占的组成成分。混淆矩阵是分类精度评价的一个标准，也是计算总体分类精度和 kappa 系数的基础。

总体分类精度（OA）等于被正确分类的像元总数除以总像元数，即

$$A_{\mathrm{OA}} = \sum_{i=1}^{k} a_{ii} / N \tag{21}$$

式中，N 为像元总数。

kappa 系数是另一种计算分类精度的方法，它的算式为[17]

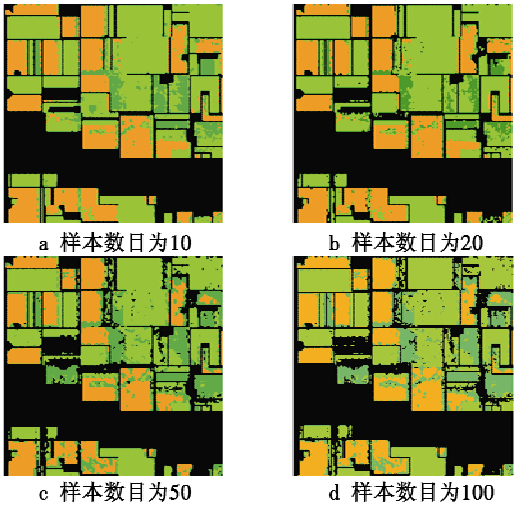
$$K = \left(N \sum_{i=1}^{k} x_{ii} - \sum_{i=1}^{k} x_{i+} x_{+i}\right) / \left(N^2 - \sum_{i=1}^{k} x_{i+} x_{+i}\right) \tag{22}$$

式中，x_{i+} 和 x_{+i} 分别表示误差矩阵第 i 行和第 i 列的元素之和。

文献[20]认为 kappa 系数的值大于 0.75 时，分类器的分类性能良好，而当 kappa 系数小于 0.4 时，性能很差。

4）分类结果与分析。

对多光谱数据 TipJul1.lan 的分类结果如图 5 所示。当标签样本的数目为 100 时，得到的混淆矩阵如表 2 所示。



a 样本数目为10　b 样本数目为20

c 样本数目为50　d 样本数目为100

图5 不同数目标签样本时的分类结果

Fig. 5 The classification with different numbers of labeled samples

表2 数据 TipJul1.lan 检测区域的混淆矩阵（数目为 100）

Table 2 The confusion matrix of the test area of TipJul1.lan

	Soybean	Corn	Alfalfa/Oats
Soybean	6692	284	247
Corn	269	8375	281
Alfalfa/Oats	628	778	1756

根据式(21)和式(22)可以计算出不同样本数目时的总体分类精度和 kappa 系数，当样本数目为 100 时得到的结果分别为 0.8712 和 0.7881。

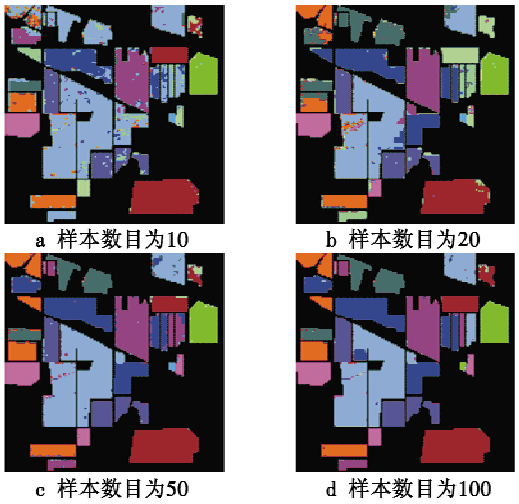
同样方法可以计算得到样本数目分别为 10，20 和 50 时的混淆矩阵，进一步可以求出各种情况下的总体分类精度和 kappa 系数，如表 3 所示。文献[21]提出的基于图的半监督分类方法的总体分类精度和 kappa 的系数分别为 0.7478 和 0.6646；使用 ISODATA 聚类算法（非监督分类）[3] 的总体分类精度和 kappa 系数分别为 0.6893 和 0.6542；文献[19]提出了使用 DBFE 特征的 ECHO 分类器是一种高质量的监督分类方法，得到的总体分类精度和 kappa 系数分别为 0.8949 和 0.8593。

表 3　实验数据 TipJul1. lan 的分类精度

Table 3　The classification accuracy for the dataset TipJul1. lan

	Number			
	10	20	50	100
Soybean	0.8011	0.8249	0.8377	0.8818
Corn	0.8143	0.8374	0.8441	0.8875
Alfalfa/Oats	0.6707	0.7132	0.7573	0.7688
OA	0.7921	0.8178	0.8313	0.8712
KC	0.7165	0.7398	0.7528	0.7881

对高光谱遥感数据 92AV3C. tif 的分类结果如图 6 所示。当标签样本的数目为 100 时，得到的混淆矩阵如表 4 所示，其中，C_i 代表第 i 类。



a 样本数目为10　　b 样本数目为20

c 样本数目为50　　d 样本数目为100

图 6　不同数目标签样本时数据 92AV3C. tif 的分类结果

Fig. 6　The classification with different numbers of labeled samples for the dataset 92AV3C. tif

表 4　数据 92AV3C. tif 检测区域的混淆矩阵(数目为 100)

Table 4　The confusion matrix of test area of dataset 92AV3C. tif

	C_2	C_3	C_5	C_6	C_8	C_{10}	C_{11}	C_{12}	C_{14}
C_2	1308	19	0	0	0	5	87	5	0
C_3	23	768	6	10	0	0	93	24	0
C_5	1	0	440	0	0	3	5	0	13
C_6	0	0	0	720	25	0	0	0	0
C_8	2	0	0	2	458	2	7	0	0
C_{10}	0	0	0	0	6	839	76	0	0
C_{11}	20	4	0	15	0	117	2097	3	2
C_{12}	78	41	5	0	0	0	102	561	47
C_{14}	2	2	48	0	0	2	1	54	1232

根据式(21)和式(22)可以计算出总体分类精度和 kappa 系数，样本数目为 100 时得到的结果分别为 0.8994 和 0.8825。

同样方法可以计算得到样本数目分别为 10，20 和 50 时的混淆矩阵，进一步可以求出各种情况下的总体分类精度和 kappa 系数，如表 5 所示。使用文献[21]提出的基于图的半监督分类方法的总体分类精度和 kappa 的系数分别为 0.747 8 和 0.664 6；使用 ISODATA 聚类(非监督分类)算法的总体分类精度和 kappa 系数分别为 0.689 3 和 0.654 2；使用 ECHO 分类器得到的分类精度和 kappa 系数分别为 0.939 6 和 0.929 5。

表 5　实验数据 2 的分类精度

Table 5　The classification results of the dataset 92AV3C. tif

	Number			
	10	20	50	100
C_2	0.7894	0.8345	0.8705	0.9121
C_3	0.7765	0.8438	0.8897	0.9209
C_5	0.8330	0.8472	0.8658	0.8853
C_6	0.8808	0.9075	0.9472	0.9639
C_8	0.8571	0.8683	0.8807	0.8998
C_{10}	0.8116	0.8432	0.8771	0.8867
C_{11}	0.7490	0.7691	0.8092	0.8479
C_{12}	0.8251	0.8667	0.8911	0.9137
C_{14}	0.8740	0.8936	0.9294	0.9521
OA	0.8071	0.8375	0.8726	0.8994
KC	0.7551	0.8137	0.8523	0.8825

一个比较自然的假设是，随着标签样本数目的增加，总体分类精度和 kappa 系数都有一定程度的提高，在标签样本较小时也有优于非监督分类的较好的分类效果。从视觉效果来看，本文方法分类之后的边界不很明显也不完整。

将本文方法分别应用于多光谱和高光谱图像，取得了较好的分类结果，并与非监督分类和监督分类方法做了比较，本文方法的分类精度与非监督分类方法相比有较大的提高，kappa 系数也更高，说明本文方法具有较高的分类精度和更好的性能，但与当前高级的监督分类方法相比仍有些差距。

4　结论

本文提出一种基于聚类集成的半监督多/高光谱图像分类方法，利用谱聚类算法的优点，将谱聚类作为聚类成员构造聚类集成系统，使用高斯 RBF 核映射下的多维数据的光谱角制图计算权值矩阵 $\boldsymbol{W}$，并用 Nyström 方法来降低方法的运算复杂度，在理论上证明了可以减少运算复杂度的结论，由于不采用 Nyström 方法而直接应用基于聚类集成的半监督分类方法在 Matlab 中运算时会提醒“Out of memory”，所以在实际实验中并没有验证这个结论。最后通过实验验证了本文方法无论对多光谱还是高光谱都有较好的分类效果，而且分类效果优于非监督分类，与高级的监督分类

方法相比稍差。此外,标签样本的选取技巧是本文没有涉及的领域,有些情况下,标签样本的获取是困难的且只能获取有限的标签样本,这种情况下应用半监督分类不涉及标签样本的选取,而有些处于应用需求的实现选择标签先验知识,则要求对标签样本选取的技巧,这也是未来要继续研究的工作。

参 考 文 献

[1] MATHER P M, KOCH M. Computer processing of remotely-sensed images [M]. 4th ed. New York: A John Wiley & Sons, Ltd., Publication, 2011:229-284.

[2] ALAJLAN N, BAZI Y, MELGANI F, et al. Fusion of supervised and unsupervised learning for improved classification of hyperspectral images [J]. Information Sciences, 2012, 217:39-55.

[3] 罗小波,赵春晖,潘建平,等. 遥感图像智能分类及其应用[M]. 北京:电子工业出版社,2011:55-57. (LUO X B, ZHAO C H, PAN J P. et al. Remote sensing image intelligent classification and its application [M]. Beijing: Publishing House of Electronics Industry, 2011:55-57.)

[4] JOHNSON B, XIE Z X. Unsupervised image segmentation evaluation and refinement using a multi-scale approach[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2011, 66(4):473-483.

[5] DUNDAR M M, LANDGREBE D A. A cost-effective semi-supervised classifier approach with kernels [J]. IEEE Transactions on Geoscience and Remote Sensing, 2004, 42 (1):264-270.

[6] 王娜,李霞. 基于监督信息特性的主动半监督谱聚类算法[J]. 电子学报,2010,38(1):172-176. (WANG N, LI X. Active semi-supervised spectral clustering based on pairwise constraints [J]. Acta Electronica Sinica, 2010, 38(1): 172-176.)

[7] 王雪松,张晓丽,程玉虎,等. 一种基于谱聚类的聚类核半监督支持向量机[J]. 中国矿业大学学报,2010,39(6):886-890. (WANG X S, ZHANG X L, CHENG Y H, et al. A cluster kernel semi-supervised support vector machine based on spectral clustering[J]. Journal of China University of Mining & Technology, 2010, 39(6):886-890.)

[8] MANTRACHA A, VAN ZEEBROECK N, FRANCQ P, et al. Semi-supervised classification and betweenness computation on large, sparse, directed graphs [J]. Pattern Recognition, 2011, 44:1212-1224.

[9] CAMPS-VALLS G, BANDOS T, ZHOU D Y. Semi-supervised graph-based hyperspectral image classification[J]. IEEE Transactions on Geoscience and Remote Sensing, 2007, 45(10):3044-3054.

[10] 刘汉强,赵凤. 基于空间特征的谱聚类含噪图像分割[J]. 模式识别与人工智能,2012,25(3):419-425. (LIU H Q, ZHAO F. Space feature based spectral clustering for noisy image segmentation[J]. PR & AI, 2012, 25(3):419-425.)

[11] 马秀丽,焦李成. 基于分水岭-谱聚类的SAR图像分割[J]. 红外与毫米波学报,2008,27(6):452-456. (MA X L, JIAO L C. SAR image segmentati on based on watershed and spectral clustering [J]. J. Infrared Millim. Waves, 2008, 27(6):452-456.)

[12] REBAGLIATI N, VERRI A. Spectral clustering with more than K eigenvectors [J]. Neurocomputing, 2011, 74(9): 1391-1401.

[13] 周林,平西建,徐森,等. 基于谱聚类的聚类集成算法[J]. 自动化学报,2012,38(8):1335-1342. (ZHOU L, PING X J, XU S, et al. Cluster ensemble based on spectral clustering[J]. Acta Automatica Sinica, 2012, 38 (8):1335-1342.)

[14] STREHL A, GHOSH J. Cluster ensembles-a knowledge reuse framework for combining partitions[J]. Journal of Machine Learning Research, 2002, 3:583-617.

[15] TUMER K, AGOGINO A K. Ensemble clustering with voting active clusters [J]. Pattern Recognition Letters, 2008, 29(14):1947-1953.

[16] MIMAROGLU S, ERDIL E. Combining multiple clusterings using similarity graph [J]. Pattern Recognition, 2011, 44:694-703.

[17] YU Z W, WONG H S, YOU J, et al. Hybrid cluster ensemble framework based on the random combination of data transformation operators [J]. Pattern Recognition, 2012, 45:1826-1837.

[18] VANDER MEER F. The effectiveness of spectral similarity measures for the analysis of hyperspectral imagery [J]. International Journal of Applied Earth Observation and Geoinformation, 2006, 8(1):3-17.

[19] LANDGREBE D. Hyperspectral image data analysis[J]. Signal Processing Magazine, IEEE, 2002, 19(1):17-28.

[20] MONTSERUD R A, LEAMANS R. Comparing global vegetation maps with the kappa statistic [J]. Ecological Modeling, 1992, 62(4):275-293.

[21] 高恒振. 高光谱遥感图像分类技术研究[D]. 长沙:国防科学技术大学,2011. (GAO H Z. Research on classification technique for hyperspectral remote sensing imagery[D]. Changsha: National University of Defense Technology, 2011.)