改进的自适应谱聚类 NJW 算法

李金泽 徐喜荣 潘子琦 李晓杰
(大连理工大学计算机科学与技术学院 大连 116024)

摘 要 聚类算法是近年来国际上机器学习领域的一个新的研究热点。为了能在任意形状的样本空间上聚类,学者们提出了谱聚类和图论聚类等优秀的算法。首先介绍了图论聚类算法中的谱聚类经典 NJW 算法和 NeiMu 图论聚类算法的基本思路,提出了改进的自适应谱聚类 NJW 算法。提出的自适应 NJW 算法的优点在于无需调试参数,即可自动求出聚类个数,克服了经典 NJW 算法需要事先设置聚类个数且需反复调试参数 δ 才能得出数据分类结果的缺点。在 UCI 标准数据集及实测数据集上对自适应 NJW 算法与经典 NJW 算法、自适应 NJW 算法与 NeiMu 图论聚类算法进行了比较。实验结果表明,自适应 NJW 算法方便快捷,且具有较好的实用性。

关键词 谱聚类,图论聚类,本征间隙
中图法分类号 TP391 **文献标识码** A

Improved Adaptive Spectral Clustering NJW Algorithm

LI Jin-ze XU Xi-rong PAN Zi-qi LI Xiao-jie
(School of Computer Science and Technology, Dalian University of Technology, Dalian 116024, China)

Abstract Clustering algorithm is a new research hotspot in the field of machine learning in recent years. In order to cluster in the sample space of any shape, the scholars have put forward excellent algorithms such as spectral clustering and graph theory clustering. In this paper, we first introduced the basic idea of the classical NJW algorithm of graph theory clustering algorithm and NeiMu algorithm, and then we gave a new algorithm named Improved adaptive spectral clustering NJW algorithm. This approach overcomes the shortcomings of the classical NJW algorithm, which needs to adjust the number of clusters in advance and needs to be repeated to obtain the data classification results. We compared the adaptive NJW algorithm with the classical NJW algorithm, the adaptive NJW algorithm and the NeiMu graph theory clustering algorithm on the UCI standard data set and the measured data set. The experimental results show that the adaptive NJW algorithm is convenient and well practicable.

Keywords Spectral clustering, Graph theory clustering, Intrinsic gap

1 引言

聚类分析是机器学习领域中的一个重要分支[1],是人们认识和探索事物之间内在联系的有效手段。所谓聚类(clustering),就是将数据对象分组成多个类或簇(cluster),使得在同一簇中的对象之间具有较高的相似度,而不同簇中的对象之间差别较大。传统的聚类算法,如 k-means 算法、EM 算法等,都是建立在凸球形的样本空间上,当样本空间不为凸时,算法会陷入局部最优。

为了能在任意形状的样本空间上聚类,学者们提出了谱聚类和图论聚类等优秀的算法。国内的张亚平等人[2]发表了一篇基于密度敏感的自适应谱聚类算法,提出了一种基于密度敏感的相似性度量,有效地描述了数据的实际聚类分布。王娜等人[3]提出了一种主动式半监督的谱聚类算法,使得类内各点紧凑而类间散布。谱聚类经典算法即 NJW 算法可以有效地对数据进行聚类分类,但其相似性矩阵的参数需大量人工调试方可确定,并且传统的 NJW 算法的聚类个数需人为规定,不能自动计算。为此,Zelnik-Manor 和 Perona[18]提出了 Self-Tuning 谱聚类算法,他们利用一种称为"Local Scaling"的思想很好地解决了相似性矩阵的参数问题。国内的孔万增等人[4]提出了基于本征间隙与正交特征向量的自动谱聚类,解决了经典谱聚类 NJW 算法需要人为输入聚类个数的问题。基于图论的聚类算法 NeiMu(Neighboring Mutually)[5]首先分析数据中的对象,寻找每个对象的 k-近邻,根据 k-近邻关系构造 k-近邻有向图,然后通过 k-近邻有向图中的 k-互邻居关系构造 k-聚类图,从而发现数据中的自然聚类。

2 图论聚类的基本理论

图论聚类法最早是由 Zahn 提出来的,又称作最大(小)支撑树聚类算法[6]。图论聚类法的前提是要把聚类的数据表示

本文受国家大学生创新创业训练项目(2016101410168)资助。
李金泽(1995-),男,主要研究方向为深度学习、神经网络,E-mail:1012638836@qq.com;徐喜荣(1967-),女,博士,副教授,博士生导师,主要研究方向为互连网络拓扑结构理论及应用、计算机算法设计与分析;李晓杰(1995-),男,主要研究方向为人工智能、模式识别。

成一个带权的无向图。图是由一些节点以及连接节点之间的边或线所构成的几何图形。在图论中，它可以反映一些对象之间的关系。两点之间不带箭头的连线称为无向边。由点和无向边构成的图叫无向图，记作 $G=(V,E)$，其中 V 是图 G 的点集合，E 是图 G 的边集合。对于无向图 G 的每一条边 (v_i, v_j)，相应地有一个数 w 表示该边上的权。权函数一般定义为两节点之间的相似度[7]。

谱聚类算法最早是从谱图划分理论演化而来。设将每一个样本数据看成是图中的顶点 V，根据样本两两之间的相似度为相应顶点之间的连接边 E 赋权重 W，于是就得到基于样本的相似度的无向加权图 $G=(V,E,W)$。从图论的最优划分理论来看，类似于线性判别算法，就是使划分成的若干子图之间相似度最小，子图内部相似度最大[8]。常用的划分准则有最小割集、规范割集、平均割集、比例割集以及最小最大割集等准则[9-11]。图划分问题的最优解是一个 NP 问题，一个较好的求解方法是考虑问题的连续松弛形式，于是便可将原问题转换成相似度矩阵或 Laplacian 矩阵的谱分解。

2.1　经典谱聚类 NJW 算法

根据不同的准则函数及谱映射方法，谱聚类算法发展了很多不同的具体实现方法，但是都可以归纳为下面 3 个主要步骤[12]：1）构建表示样本集的矩阵 Z；2）通过计算矩阵 Z 的前 k 个特征值与特征向量，构建特征向量空间；3）利用 k-means 或其他经典聚类算法对特征向量空间中的特征向量进行聚类。

经典谱聚类 NJW 算法的主要步骤如下：

Step 1　构建相似矩阵 A。矩阵 A 中的元素 $A_{ij}=\exp(-\frac{d^2(v_i,v_j)}{2\delta^2})$，其中，$v_i$ 是数据样本点，$d(v_i,v_j)$ 代表两样本点之间的欧氏距离 $\| v_i-v_j \|$，σ 是决定样本点之间衰减速度的尺度参数。

Step 2　构造规范性 Laplacian 矩阵 $Lsym$。首先构造度矩阵 D，度矩阵 D 的主对角线上的元素 $D(i,j)$ 为相似性矩阵 A 的第 i 行元素之和，其他元素均为 0；然后构造拉普拉斯矩阵 $L=D-A$ 为 $L=D^{-\frac{1}{2}}AD^{-\frac{1}{2}}$；进一步再构造规范性 Laplacian 矩阵 $Lsym=D^{-\frac{1}{2}}LD^{-\frac{1}{2}}=E-D^{-\frac{1}{2}}AD^{-\frac{1}{2}}$。

Step 3　计算矩阵 Lsym 的前 k 个最大特征值对应的特征向量 $x_1,\cdots,x_k$（必要时需做正交化处理），此 k 值需人为输入，并据此构造矩阵 $X=[x_1,\cdots,x_k]\in Rn\times k$，将矩阵 X 的行向量转变为单位向量，得到矩阵 Y：$Y_{ij}=\frac{X_{ij}}{\sqrt{\sum_j X_{ij}^2}}$。将矩阵 Y 的每一行看作是空间中的一个点，对其使用 k 均值算法或任意其他经典聚类算法，得到 k 个聚类；将数据点 y_i 划分到聚类 j 中，当且仅当 Y 的第 i 行被划分到聚类 j 中。

2.2　NeiMu 图论聚类算法

基于 NeiMu(Neighboring Mutually)图论的聚类算法首先分析数据中的对象，寻找每个对象的 k-近邻，根据 k-近邻关系构造 k-近邻有向图，然后通过 k-近邻有向图中的 k-互邻居关系构造 k-聚类图，发现数据中的自然聚类。该算法的特点是根据数据之间的互为 k-近邻关系确定数据中的自然簇，而不必引入其他方法来划分小簇，从而能够保证对象不会被错误聚类，仅会与其他小簇一起融合到一个大簇中。这一优点可以有效保证 NeiMu 算法的聚类质量。NeiMu 算法的步骤如下所示[4]：

Step 1　构造距离矩阵 D

$$D=\begin{bmatrix} 0 & d(1,2) & \cdots & d(1,n) \\ d(2,1) & 0 & \cdots & d(2,1) \\ \vdots & \vdots & \ddots & \vdots \\ d(n,1) & d(n,2)\ ) & \cdots & 0 \end{bmatrix} \tag{1}$$

其中，$d(i,j)$ 为对象 i 和对象 j 之间的距离，满足 $d(i,j)\geqslant 0$，$d(i,j)=d(j,i)$ 且 $d(i,j)=0$。在计算和存储中，k-近邻有向图 $kNN\_G=(W,E)$ 可以表示为邻接矩阵的形式，矩阵的元素 $A_{KNN_G}(i,j)$ 如式(2)所示：

$$A_{KNN_G}(i,j)=\begin{cases} 1, & \text{如果边 } W_j \text{ 是 } W_i \text{ 的 k-近邻} \\ 0, & \text{否则} \end{cases} \tag{2}$$

Step 2　构造 k-聚类图

$$A_{KNN_G}(i,j)=\begin{cases} 1, & \text{如果 } A_{KNN_G}(i,j)*A_{KNN_G}(j,i)=1 \\ 0, & \text{如果 } A_{KNN_G}(i,j)*A_{KNN_G}(j,i)=0 \end{cases} \tag{3}$$

Step 3　遍历 k-聚类图

k-聚类图包含若干个由 k-互邻居组成的连通子图（当子图仅包含一个结点时，子图退化成孤立点）。因此只需对 k-聚类图进行深度优先搜索遍历，即可得到最终的聚类结果，使得 k-聚类图中每个连通子图的遍历结果构成一个聚类。

3　改进的自适应谱聚类 NJW 算法

传统谱聚类 NJW 算法能比较准确地对数据进行聚类，但仍有很多不足之处。在该算法的 Step 1 中，高斯核函数的尺度参数 σ 需要人为规定，并且需反复调试方可确定其最优值。在 Step 3 中，需要人为规定聚类个数 k，无法自动获得聚类的个数。为此，Zelnik-Manor 和 Perona[18] 提出了 Self-Tuning 谱聚类算法，他们利用一种被称为“Local Scaling”的思想很好地解决了 Step 1 中函数的参数 σ 无法准确确定的问题。国内的卜德云[13] 在 Self-Tuning 的基础上做了关于自适应谱聚类算法的相关研究。

本文采取 Self-Tuning 的相关算法对经典谱聚类 NJW 算法中的 Step 1 进行改进，将 Step 1 中的 $A_{ij}=\exp(\frac{-d^2(v_i,v_j)}{2\delta^2})$ 改进为 $A_{ij}=\exp(\frac{-d^2(s_i,s_j)}{\sigma_i\sigma_j})$，其中 $\sigma_i=d(s_i,s_T)$ 是样本点 i 到其第 T 个近邻的点的欧氏距离。文献[14-16]测试了多个人工数据集和实际数据集，认为 T 取 7 时算法的效果最佳，这里不做多余讨论。

对经典谱聚类 NJW 算法中 Step 3 采取了基于本征间隙的自动谱聚类算法，解决了传统 NJW 算法无法自动计算聚类个数的问题。计算规范 Laplacian 矩阵的特征值，并按顺序从大到小排列为 $\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_n$，计算本征间隙序列 $\{g_1, g_2,\cdots,g_{n-1}\cdot g_i=\lambda_i-\lambda_{i+1}\}$。根据矩阵的扰动理论，在本征间隙序列中依次寻找第一个极大值，则该值对应的下标即为类个数[19]，即类别数 $k=\arg\min\{g_i-g_{j|j<i}>0\ \&\&\ g_i-g_{i+1}>0\}$[17]。

改进的自适应谱聚类 NJW 算法的步骤如下：

Step 1 构建相似矩阵 A。矩阵 A 中的元素 $A_{ij}=\exp(\frac{-d^2(s_i,s_j)}{\sigma_i\sigma_j})$，其中 $\sigma_i=d(s_i,s_T)$，是样本点到其第 T 个近邻点的欧氏距离，v_i 是数据样本点，$d(v_i,v_j)$代表两样本点之间的欧氏距离 $\|v_i-v_j\|$。

Step 2 首先构造度矩阵 D，度矩阵 D 的主对角线上的元素 $D(i,j)$为相似性矩阵 A 的第 i 行元素之和，其他元素均为 0；再构造拉普拉斯矩阵 $L=D-A$ 为 $L=D^{-\frac{1}{2}}AD^{-\frac{1}{2}}$；然后进一步计算规范性 Laplacian 矩阵 $Lsym=D^{-\frac{1}{2}}LD^{-\frac{1}{2}}=E-D^{-\frac{1}{2}}AD^{-\frac{1}{2}}$。

Step 3 计算规范 Laplacian 矩阵的特征值，并按顺序从大到小排列为 $\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_n$，计算本征间隙序列$\{g_1,g_2,\cdots,g_{n-1}\mid g_i=\lambda_i-\lambda_{i+1}\}$，在本征间隙序列中依次寻找第一个极大值，则该值对应的下标即为类个数 k；计算矩阵 Lsym 的前 k 个最大特征值对应的特征向量 $x_1,\cdots,x_k$（必要时需做正交化处理）。据此构造矩阵 $X=[x_1,\cdots,x_k]\in Rn\times k$，将矩阵 X 的行向量转变为单位向量，得到矩阵 Y：$Y_{ij}=\frac{X_{ij}}{\sqrt{\sum_j X_{ij}^2}}$，将矩阵 Y 的每一行看作是 R_k 空间中的一个点，对其使用 k-均值算法或任意其他经典聚类算法得到 k 个聚类；将数据点 y_i 划分到聚类 j 中，当且仅当 Y 的第 i 行被划分到聚类 j 中。

4 实验分析与比较

谱聚类是基于谱图划分理论的一种聚类算法，是图论聚类的一个分支。自适应 NJW 算法是对经典谱聚类 NJW 算法的一种改进。为了得出改进的谱聚类自适应 NJW 算法与经典的谱聚类 NJW 算法以及图论聚类 NeiMu 算法之间的区别，下面分别在 UCI 标准数据以及实测数据中进行了相关的比较实验。

4.1 基于 UCI 标准数据的自适应 NJW 与经典 NJW 聚类的比较

本实验在国际通用数据库 UCI 数据库（专门用于测试分类及聚类算法）中的 Tea，Iris 以及 Ionesphere 3 个数据集上进行了改进的自适应 NJW 算法与经典 NJW 算法的比较实验。本次实验中经典 NJW 算法参数 δ 的调试步长为 0.01，参与比较的自适应 NJW 算法的参数 T 取经验值 7。本次实验在计算机内存 4.0GB、CPU 3.2GHz 的硬件条件以及 Matlab 2016b 环境下运行。

本次实验中的经典 NJW 算法参数的调整步长为 0.01，k 值取数据集的固有聚类个数，准确率平均值是指参数 δ 在较好聚类效果区间内以 0.01 为步长所得到的所有准确率的平均值。实验结果如表 1 所列，从中可以看出，自适应 NJW 算法在数据分类准确率上略有提高。自适应 NJW 算法最大的优点在于无需调试参数，只需将参数 T 设置为经验值 7 即可自动且较为准确地求出了聚类个数；而经典 NJW 算法不仅需要事先设置聚类个数，且参数 δ 也很大程度上影响了数据分类准确类，需反复调试 δ 才能得出最好的数据分类结果。

表 1 经典 NJW 算法与自适应 NJW 算法的比较

	δ 的取值范围	聚类效果	经典 NJW 算法	自适应 NJW 算法
Iris	$0.16<\delta<0.25$	较好	$k=3$ 时，δ 在 0.16～0.25 得到的	无需调试参数
	$\delta<0.16$ 或 $\delta>0.25$	较差	准确率平均值为 0.903	聚类准确率为 0.920
Ionesphere	$0.15<\delta<0.25$	较好	$k=2$ 时，δ 在 0.15～0.25 得到的	无需调节参数
	$\delta<0.15$ 或 $\delta>0.3$	较差	准确率平均值为 0.721	聚类准确率为 0.746
Tea	$0.08<\delta<0.11$	较好	$k=3$ 时，δ 在 0.08～0.11 得到的	无需调试参数
	$\delta<0.08$ 或 $\delta>0.11$	较差	准确率平均值为 0.384	聚类准确率为 0.424

4.2 基于 UCI 标准数据的自适应 NJW 与图论聚类 NeiMu 算法的比较

本实验同样在 UCI 数据库中的 Iris，Ionesphere 及 Tea 3 个数据集上对谱聚类算法与图论聚类算法进行了比较。本次实验在计算机内存 4.0GB、CPU 3.2GHz 的硬件条件以及 Matlab 2016b 环境下进行。

实验结果如表 2 所列。

表 2 图论聚类算法与谱聚类算法在 UCI 数据集上的比较

	数据属性			算法性能（分类准确率）	
	样本数	维数	固有类数	NeiMu 算法（k 值）	自适应 NJW 算法
Iris	150	4	3	0.673（$k=16$）	0.920
Ionesphere	351	34	2	0.644（$k=24$）	0.746
Tea	151	5	3	0.338（$k=8$）	0.424

从表 2 可以看出，多维数据用谱聚类的自适应 NJW 算法比图论聚类 NeiMu 算法的效果好，NeiMu 算法出现了严重的分类失误，具体体现在个别簇包含过多的点，而某些簇包含的点极少，这就体现出图论聚类的不足。基于图论聚类的 NeiMu 算法对点间距离较大或者密度比较稀疏的数据集能起到很好的聚类效果，对于点间距离较小或者数据分布与点间距离没有关系的数据集的分类效果较差。

4.3 基于本征间隙的自适应谱聚类算法与 Self-Tuning 算法的比较

本实验也基于 UCI 数据库中的 Iris、Ionesphere、Tea 3 个数据集对谱聚类算法与图论聚类算法进行比较。本次实验在计算机内存 4.0GB、CPU 3.2GHz 的硬件条件以及 Matlab 2016b 环境下进行。

实验结果如表 3 所列。

表 3 基于本征间隙的谱聚类算法与 Self-Tuning 算法的比较

	数据属性			算法性能（分类准确率）	
	样本数	维数	固有类数	基于本征间隙的谱聚类算法（自动确定）	Self-Tuning 算法（预先指定）
Iris	150	4	3	0.920（$\delta=0.17$）	0.920
Ionesphere	351	34	2	0.721（$\delta=0.20$）	0.746
Tea	151	5	3	0.411（$\delta=0.10$）	0.424

从表 3 可以看出，Self-Tuning 算法是一种非自适应性算法，在对数据进行聚类分析之前需要知道数据的聚类个数，这也是 Self-Tuning 算法的不足之处，其聚类准确率原则上与本文提出的自适应 NJW 算法一致。从表 3 看出，基于本征间隙的谱聚类算法是一种自适应算法，可以根据矩阵的扰动理论[19]计算出数据集的本征间隙，并自动计算出数据的聚类个

数。但其缺点也是显而易见的，即参数 δ 的值难以确定，需反复调试。通过这两种算法的比较，择取两者的优点，本文将这两种算法应用在了经典 NJW 算法中。

4.4　基于实测数据的自适应 NJW、经典 NJW 算法与 NeiMu 图论聚类之间的比较

高场不对称波形离子迁移谱（High-field Asymmetric Waveform Ion Mobility Spectrometry，FAIMS）是一种大气环境下利用不同离子在强电场中离子迁移率非线性变化的特点来实现离子空间的分离和识别的检测技术。FAIMS 利用不同离子在强电场中离子迁移率非线性变化的特性来实现离子的分离，从而完成很多领域的数据采集工作。本文数据是由 FAIMS 仪器采集的，获得了三维空间坐标下的谱图数据（x，y，z），其中 x 轴：补偿电压 CV，共有 512 个采样点，从 -6V～$+6$V；y 轴：分离电压 DV，共有 26 个采样点，范围为 0%～100%，步长为 4%；z 轴：离子流强度 Ion Current，即每一幅谱图共有 $512*26=13312$ 个点（x，y，z）。

本文待处理的数据针对空间中的 26 条曲线，采用了谱图峰识别算法识别出了每条曲线上的若干个谱峰点所得到的谱峰点集合，一共 66 个三维数据点。

图 2 为 NeiMu 算法参数 $k=4$ 时的实验结果，可以看出数据被分为了 9 类，并且第 4 个聚类（从左向右）只含有一个孤点，表明图论聚类 NeiMu 算法对孤立点比较敏感。图 3 为经典 NJW 算法参数 $\sigma=0.16$，$k=9$ 时的实验结果，可以看出经典 NJW 算法与自适应 NJW 算法的聚类效果无异，但经典 NJW 算法需要反复调节参数 σ 的值，且需手动输入参数 k 的值，直到聚类效果较好为止。图 4 为自适应 NJW 算法的实验结果，可知数据点被分成了 9 类，与图 2 的不同之处在于自适应 NJW 算法将孤立点划分到其他聚类中，并且不需调试参数就可以准确得出聚类个数。

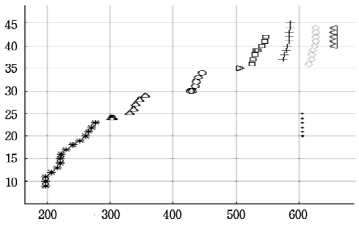

图 2　NeiMu 算法的实验结果（$k=4$）

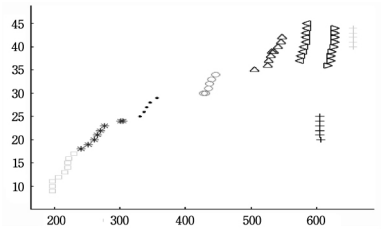

图 3　经典 NJW 算法的聚类结果（$\sigma=0.16$，$k=9$）

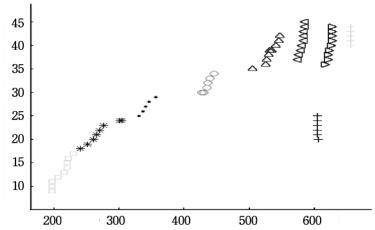

图 4　自适应 NJW 算法的聚类结果

图论聚类与谱聚类在 FAMIS 数据集上的对比如表 4 所列。

表 4　图论聚类与谱聚类在 FAMIS 数据集上的对比

	孤立点敏感	运算时间/s	参数
NeiMu	敏感	0.203	$k=4$ 最优
经典 NJW	不敏感	1.601	$\sigma=0.16$ 最优
自适应 NJW	不敏感	1.430	无需参数

由表 4 易知自适应 NJW 算法和经典 NJW 耗时比较长，且对孤立点不敏感。NJW 算法的突出优势在于其能自动确定聚类个数，不用人工调试参数，这是图论聚类 NeiMu 算法和经典 NJW 算法所不能达到的。相比图论聚类 NeiMu 算法，基于谱聚类的自适应 NJW 算法则能很好地对高维数据进行聚类，无需人工调试输入参数即可较为准确地确定聚类个数，这些是图论聚类 NeiMu 算法所不能及的。

结束语　本文将经典 NJW 算法做了两方面改进：1）在数据相似性矩阵中引入了 Self-Turning 算法思想，解决了高斯核函数参数 σ 难以确定的问题；2）采用孔万增等人提出的基于本征间隙的自动谱聚类算法，使得 NJW 算法可以较为准确地自动计算聚类个数。根据若干次改进后的自适应 NJW 算法与经典 NJW 算法和图论聚类 NeiMu 算法的对比实验，可以得出以下结论：

（1）谱聚类基于谱图的划分理论，是图论聚类的一个分支。自适应 NJW 算法是基于谱聚类的、对经典 NJW 算法的改进。自适应 NJW 算法无需调试参数的优点，使得其能方便快捷地适用于现场检测。

（2）图论聚类适用于密度变化大、大小相差大、维数较低的数据集，尤其在二维数据处理上能快速、准确地对数据进行聚类。自适应谱聚类 NJW 算法适用于维数较多、数据量较大的数据集。

（3）从实测数据上可以看出，图论聚类对孤立点敏感，谱聚类对孤立点不敏感。在对数据进行定性分析时，谱聚类优于图论聚类。在对特殊点处理时，图论聚类优于谱聚类。

参 考 文 献

[1] JAIN A，MURTY M，FLYNN P. Data clustering：A Review [J]. ACM Computing Surveys，1999，31(3)：264-323.

[2] 张亚平，杨明. 一种基于密度敏感的自适应谱聚类算法[J]. 数学的实践与认识，2013，43(20)：150-156.

[3] 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.

[4] 孔万增，孙志海，杨灿，等. 基于本征间隙与正交特征向量的自动谱聚类[J]. 电子学报，2010，38(8)：1880-1891.

[5] 应德全，应晓敏，叶继华. 一种基于图论的聚类算法 NeiMu[J]. 计算机工程与应用，2009，45(3)：47-50.

[6] 钱云涛，赵荣椿，谢维信. 鲁棒聚类—基于图论和目标函数的方法[J]. 电子学报，1998，26(2)：91-94.

[7] 刘锁兰，王江涛，王建国，等. 一种新的基于图论聚类的分割算法[J]. 计算机科学，2008，35(9)：245-247.

[8] SHI J，MALIK J. Normalized cuts and image segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2000，22(8)：888-905.

[9] 蔡晓妍，戴冠中，杨黎斌. 谱聚类算法综述[J]. 计算机科学，2008，35(7)：14-18.

（下转第 450 页）

通过图 5 及表 4 可以看出，对高斯过程采用 PSO 算法优化之后，查准率、查全率和微平均 3 项指标均有明显的提高（存在极个别偏低的现象），其中准确率提高了近 15%，已接近 80%，表明采用 PSO 进行参数优化对分类器分类效果的提升有很重要的作用。

结束语 本文实现了基于粒子群优化算法（PSO）优化高斯过程（Guassian Process）的中文文本分类。通过爬虫技术进行语料采集，利用情感词典对中文文本语料进行特征提取，计算特征权重，从而数字化表示特征，最终利用所提分类模型进行中文文本情感分类。实验结果表明，相较于单纯高斯过程分类方法，本文提出的方法较大程度地提高了其性能。

参 考 文 献

[1] 王素格，李德玉，魏英杰. 基于赋权粗糙隶属度的文本情感分类方法[J]. 计算机研究与发展，2011，48(5)：855-861.

[2] 马晓玲，金碧漪，范并思. 中文文本情感倾向分析研究[J]. 情报资料工作，2013(1)：52-56.

[3] WANG J X，DONG A. A comparison of two text representations for sentiment analysis[C]// 2010 International Conference on Computer Application and System Modeling (ICCASM 2010). IEEE，2010.

[4] 徐健锋，许园，许元辰，等. 基于语义理解和机器学习的混合的中文文本情感分类算法框架[J]. 计算机科学，2015，42(6)：61-66.

[5] 万源. 基于语义统计分析的网络舆情挖掘技术研究[D]. 武汉：武汉理工大学，2012.

[6] CASALE S，RUSSO A，SCEBBA G，et al. Speech Emotion Classification using Machine Learning Algorithms[C]// The IEEE International Conference on Semantic Computing. Santa Clara，California，USA，August 2008：4-7.

[7] CHANGLI Z，WANLI Z，TAO P，et al. Sentiment Classification for Chinese Reviews Using Machine Learning Methods Based on String Kernel[C]// Third International Conference on Convergence and Hybrid Information Technology. Pusan，Korea，November 2008：11-13.

[8] 王维博. 粒子群优化算法研究及其应用[D]. 成都：西南交通大学，2012.

[9] DAHIWALE P，RAGHUWANSHI M M，MALIK L. Design of Improved Focused Web Crawler by Analyzing Semantic Nature of URL and Anchor Text[C]// 2014 9th International Conference on Industrial and Information Systems. Gwalior，India，December 2014：15-17.

[10] W3C Document Object Model Level 3 Core Specification[OL]. http://www.w3.org/TR/DOM-Level-3-Core.

[11] 湛燕，陈昊，袁方，等. 基于中文文本分类的分词方法研究[J]. 计算机工程与应用，2003，29(23)：87-88，91.

[12] 高丹，张彦霞，赵永恒. 中国虚拟天文台交叉证认工具的开发和应用[J]. 天文学报，2008，49(3)：348-358.

[13] 徐琳宏，林鸿飞，杨志豪. 基于语义理解的文本倾向性识别机制[J]. 中文信息学报，2007，21(1)：96-100.

[14] 吴艳玲. 基于 SVM 的网页分类器的研究[D]. 长春：吉林大学，2004.

[15] 石慧. 基于特征选择和特征加权算法的文本分类研究[D]. 济南：山东师范大学，2015.

[16] 王之鹏. Web 文本分类系统中文本预处理技术的研究与实现[D]. 南京：南京理工大学，2009.

[17] 平源. 基于支持向量机的聚类及文本分类研究[D]. 北京：北京邮电大学，2012.

[18] PLATANIOS E A，CHATZIS S P. Gaussian Process-Mixture Conditional Heteroscedasticity[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2014，36(5)：888-900.

[19] 胡明. 基于高斯过程的非线性预测控制[D]. 广州：华南理工大学，2012.

[20] 王洪春. 贝叶斯公式与贝叶斯统计[J]. 重庆科技学院学报（自然科学版），2010，12(3)：203-205.

[21] 王成，刘亚峰，王新成，等. 分类器的分类性能评价指标[J]. 电子设计工程，2011，19(8)：13-15，21.

（上接第 427 页）

[10] FILIPPONEA M，CAMASTRAB F，MASULLIA F，et al. A survey of kernel and spectral methods for clustering[J]. Pattern Recognition，2008，41(1)：176-190.

[11] LUXBURG U. A tutorial on spectral clustering[J]. Statistics and Computing，2007，17(4)：395-416.

[12] VERMA D，MEILA M. A comparison of spectral clustering algorithms：2003. UW CSE Tech nical report 03-05-01[R]. 2003.

[13] 卜德云，张道强. 自适应谱聚类算法研究[J]. 山东大学学报，2009，39(5)：22-39.

[14] ZHAO F，JIAO L C，LIU H Q，et al. Spectral Clustering with Eigenvector Selection Based on Entropy anking[J]. Neurocomputing，2010，73(10-12)：1704-1717.

[15] 赵凤，焦李成，刘汉强，等. 半监督谱聚类特征向量选择算法[J]. 模式识别与人工智能，2011，24(1)：48-56.

[16] ZELNIK-MANOR L，PERONA P. Self-Tuning Spectral Clustering[C]// Proc of Advances in Neural Information Processing Systems 17. Vancouver，Canada，2004：1601-1608.

[17] NGA Y，JORDAN M I，WEISS Y. On spectral clustering：Analysis and an algorithm[C]// Proceedings of the 14th Advances in Neural Information Processing Systems (NIPS 2002). Cambridge，MIT Press，2002：849-856.

[18] ZELNIK M L，PERONA P. Self-tuning spectral clustering[C]// Advances in Neural Information Processing Systems (NIPS). Cambridge，MA：MIT Press，2004.

[19] 孙继广. 矩阵扰动分析[M]. 北京：科学出版社，2001：146-160.