第36卷 第15期 计 算 机 工 程 2010年8月
Vol.36 No.15 Computer Engineering August 2010
·软件技术与数据库· 文章编号：1000—3428(2010)15—0040—03 文献标识码：A 中图分类号：TP181

用于文档聚类的间隔流形学习算法研究

李 昕，钱 旭，王自强
(中国矿业大学(北京)机电与信息工程学院，北京 100083)

摘 要：为有效解决文档聚类问题，提出一种基于间隔流形学习的文档聚类算法。该算法利用间隔Fisher分析将高维文档空间降维到低维特征空间，利用支持向量聚类算法进行聚类。在基准文档测试集上的实验结果表明，该算法的聚类性能优于其他常用的文档聚类算法。
关键词：文档聚类；流形学习；支持向量聚类；数据挖掘

Research on Marginal Manifold Learning Algorithm for Document Clustering

LI Xin, QIAN Xu, WANG Zi-qiang
(College of Mechanical Electronic and Information Engineering, China University of Mining and Technology(Beijing), Beijing 100083)

【Abstract】To effectively deal with the document clustering problem, a novel document clustering algorithm based on marginal manifold learning is proposed. The high dimensional document space is reduced into the lower dimensional feature space with marginal fisher analysis. The support vector clustering algorithm is applied to cluster documents herein. Experimental results on the benchmark document sets show the algorithm achieves much better clustering performance than tradition document clustering algorithms.
【Key words】document clustering; manifold learning; support vector clustering; data mining

1 概述

文档聚类作为一种无监督的机器学习方法，旨在根据文档的内容将文档集合划分为若干个类，以便使得每个类内的文档数据具有高度的内在相似性，而不同类间的文档数据差别较高。由于文档聚类在Web数据挖掘、自动文摘生成、搜索引擎、个性化推荐系统等领域具有广阔的应用前景，因此如何设计高效的文档聚类算法日益成为数据挖掘和机器学习研究领域的热点问题之一。

由于文档数据空间通常具有很高的维数，因此传统的聚类算法在处理高维文档数据时存在着搜索时间过长、聚类准确率不高的问题。为了有效地克服文档聚类时的“维数灾难”问题，近年来，基于维数降维的文档聚类算法日益成为研究的热点，其算法思想为：首先将高维文档数据降维到低维特征空间，然后再采用传统的聚类算法进行文档聚类。其代表性算法有：潜在语义索引法，谱聚类法[1]，非负矩阵分解法[2]和局部保持索引法[3]等。但是这些算法的不足之处在于：(1)没有显式地利用间隔思想来描述不同类间的分离性；(2)仍然采用传统的K-均值算法来实现文档聚类算法，其聚类准确率有待进一步提高。为了克服上述不足，进一步提高文档聚类算法的性能，本文提出一种基于间隔流形学习的文档聚类算法。实验结果验证了该算法的可行性和高效性。

2 研究基础

2.1 间隔流形学习

间隔Fisher分析(Marginal Fisher Analysis, MFA)[4]是一种最近提出的用于维数降维的间隔流形学习算法。该算法在数据降维的过程中，能够利用类内紧凑图和类间分离图来分别描述类内相似数据之间的紧凑性和不同类间数据的分离性，有效地克服了传统流形学习算法[5]的如下不足：仅用单一邻接图来对局部流形结构进行建模，而没有显式地利用间隔思想来描述不同类间的分离性。

下面给出MFA的算法描述：

Step1 构造类内紧凑图和类间分离图。设给定 n 个数据集 $X=\{x_1,x_2,\cdots,x_n\}$，其中，$x_i\left(\in\square^D\right)$ 是高维数据。设 G 和 G^p 分别表示2个具有 n 个顶点的无向图：类内紧凑图和类间分离图，其中，顶点 i 对应于样本数据点 x_i。

在类内紧凑图 G 中，对于任意数据点 x_i，如果 x_i 和 x_j 具有相同的类别，且满足 x_j 是 x_i 的 k_1- 最近邻居，则在数据点 x_i 和 x_j 之间创建一条边。

在类间分离图 G^p 中，对于任意数据点元组 $\left(x_i,x_j\right)$，如果满足如下2个条件：(1) x_i 是 x_j 的 k_2- 最近邻居，且 $x_i\left(\in C_1\right)$ 和 $x_j\left(\in C_2\right)$ 来自不同的类，即类 $C_1\neq C_2$；(2) x_i 和 x_j 之间的距离是所有分别来自类 C_1 和 C_2 中的数据点组成的 k_2 -最近邻居元组中距离最短的，则在 x_i 和 x_j 之间创建一条边。

Step2 确定图 G 和 G^p 中边的权重。在类内紧凑图 G 中，如果顶点 i 和 j 之间有条边，则设其边权重 $S_{ij}=1$，否则为0。在类间分离图 G^p 中，如果顶点 i 和 j 之间有条边，则设其边权重 $S_{ij}^p=1$，否则为0。

Step3 计算MFA嵌入映射。通过计算如下间隔Fisher优化函数：

基金项目：教育部科学技术研究基金资助重点项目(107021)
作者简介：李 昕(1978－)，男，博士研究生，主研方向：数据挖掘；钱 旭，教授、博士、博士生导师；王自强，博士研究生
收稿日期：2009-12-07 **E-mail：**lixincumtb@126.com

$$w_{opt}=\arg\min_{w}\frac{\boldsymbol{w}^{\mathrm{T}}\boldsymbol{X}\left(\boldsymbol{D}-\boldsymbol{S}\right)\boldsymbol{X}^{\mathrm{T}}\boldsymbol{w}}{\boldsymbol{w}^{\mathrm{T}}\boldsymbol{X}\left(\boldsymbol{D}^{p}-\boldsymbol{S}^{p}\right)\boldsymbol{X}^{\mathrm{T}}\boldsymbol{w}} \tag{1}$$

其中，$\boldsymbol{w}$ 为可获得特征向量；$\boldsymbol{D}$ 和 $\boldsymbol{D}^{p}$ 是对角矩阵，定义如下：

$$\boldsymbol{D}=\sum_{j}S_{ij},\quad \boldsymbol{D}^{p}=\sum_{j}S_{ij}^{p} \tag{2}$$

设 $\boldsymbol{w}_1,\boldsymbol{w}_2,\cdots,\boldsymbol{w}_d$ 是式(1)的特征向量，并按照其所对应的特征值由小到大的顺序排列，即 $\lambda_1<\lambda_2<\cdots<\lambda_d$，则由高维数据空间到低维特征空间的映射可表示为

$$x_i\rightarrow y_i=\boldsymbol{W}^{\mathrm{T}}x_i \tag{3}$$

其中，$\boldsymbol{W}=\left[\boldsymbol{w}_1,\boldsymbol{w}_2,\cdots,\boldsymbol{w}_d\right]$，$y_i\left(\in\square^{d}\right)$ 是高维数据 x_i 的低维嵌入。

2.2 支持向量聚类

为了准确地对降维后的文档数据进行聚类，采用基于统计学习理论的支持向量聚类(Support Vector Clustering, SVC)[6]算法。该算法的优点为：无需预先指定聚类的数目，能处理任意形状的聚类，且能对噪声及孤立点进行有效的处理。其算法过程主要包括 2 个部分：基于支持向量机训练和聚类标识。下面给出 SVC 的实现过程：

设 $\{x_i\}_{i=1}^{n}\subseteq\square$ 是一个 d－维输入数据集合，SVC 首先利用非线性变换函数 Φ 将 x_i 将映射到高维特征空间 F，然后在其中寻找一个能包含所有数据点且具有最小半径的超球体，其形式化定义如下：

$$\min\left(R^2+C\sum_{i=1}^{n}\xi_i\right) \tag{4}$$

约束条件为

$$\left\|\Phi\left(x_i\right)-a\right\|^2\leqslant R^2+\xi_i,\ \xi_i>0 \tag{5}$$

其中，R 表示能包含数据集的最小球半径；a 表示球的中心；ξ_i 表示非负松弛变量，用来允许一部分数据位于球的外面；C 表示大于零的惩罚因子。

为了求解在式(5)约束下的优化问题，可定义如下拉格朗日函数：

$$L=R^2-\sum_{i=1}^{n}\left(R^2+\xi_i-\left\|\Phi\left(x_i\right)-a\right\|^2\right)\beta_i-\sum_{i=1}^{n}\xi_i\mu_i+C\sum_{i=1}^{n}\xi_i \tag{6}$$

其中，$\beta_i\geqslant 0$ 和 $\mu_i\geqslant 0$ 是拉格朗日系数。

为了求得式(6)的极小值，令 L 对变量 R，a 和 ξ_i 的偏导数分别等于零，可得如下公式：

$$\sum_{i=1}^{n}\beta_i=1 \tag{7}$$

$$a=\sum_{i=1}^{n}\beta_i\Phi\left(x_i\right) \tag{8}$$

$$\beta_i=C-\mu_i \tag{9}$$

另外，由 KKT 补充条件可得如下结论：

$$\xi_i\mu_i=0 \tag{10}$$

$$\left(R^2+\xi_i-\left\|\Phi\left(x_i\right)-a\right\|^2\right)\beta_i=0 \tag{11}$$

通过消去变量 R，a 和 ξ_i，可将式(6)转化为如下 Wolfe 对偶形式：

$$WL=\sum_{i=1}^{n}K\left(x_i,x_i\right)\beta_i-\sum_{i,j=1}^{n}\beta_i\beta_jK\left(x_i,x_j\right) \tag{12}$$

约束条件为

$$0\leqslant\beta_i\leqslant C,\ \sum_{i=1}^{n}\beta_i=1 \tag{13}$$

其中，$K(,)$ 是满足 Mercer 定理的核函数，用来计算输入空间在特征空间的点积形式：

$$K\left(x_i,x_j\right)=\Phi\left(x_i\right)\cdot\Phi\left(x_j\right) \tag{14}$$

需要说明的是：当 $\beta_i=0$ 时，其所对应的数据点位于超球的内部；当 $\beta_i=C$ 时，其对应的数据点称为边界支持向量，位于特征空间中超球的外面；当 $0<\beta_i<C$ 时，其对应的数据点称为支持向量，位于特征空间中超球的表面上，它们描述了数据类的轮廓。

另外，输入空间的任意一点 x 在特征空间的映射点 $\Phi(x)$ 到球心的距离可表示为

$$R(x)=\left\|\Phi(x)-a\right\|^2=K(x,x)-2\sum_{i}\beta_iK\left(x_i,x\right)+\sum_{i,j}\beta_i\beta_jK\left(x_i,x_j\right) \tag{15}$$

于是，任意支持向量 $\boldsymbol{x}_i$ 到球心的距离即为球半径，即：$R=R(\boldsymbol{x}_i)$。所以，聚类的边界曲线可由满足 $\left\{x\middle|R(x)=R\right\}$ 的数据点组成的轮廓线集构成。其中，支持向量在聚类边界线上，边界支持向量在聚类边界线外，其余点在聚类边界线的内部。

设 x_i 和 x_j 是属于不同聚类的输入数据，则由以上讨论可知：在映射后的特征空间中，连接它们的任意路径上必然存在数据点 z 满足：$R(z)>R$。因此，确定聚类的内容时，首先计算如下邻接矩阵 $\boldsymbol{A}$：

$$A_{ij}=\begin{cases}1 & \text{如果对于}x_i\text{和}x_j\text{连线的所有点}z\text{都满足}R(z)<R\\ 0 & \text{其他}\end{cases} \tag{16}$$

然后根据邻接矩阵 A 确定各个连通分量，一个连通分量就是一个聚类。对于那些包含一个数据点的连通分量，则标记为噪声或者孤立点。

3 基于MFA和SVC的文档聚类算法

设给定文档数据集 $\tilde{X}=\{\tilde{x}_i\}_{i=1}^{n}$，其中，$\tilde{x}_i\in\square^{D}$，采用向量空间模型(VSM)中的 TF－IDF 项权重向量来表示每个文档 $\tilde{x}_i$，并将其作归一化处理。本文提出的文档聚类的算法思想为：首先利用间隔流形学习算法 MFA 将高维文档数据降维到低维特征空间，然后在其中利用支持向量聚类算法 SVC 进行文档聚类。其算法过程描述如下：

Step1 文档预处理。为了使间隔流形学习算法 MFA 的优化目标不包含平凡解，即式(1)的分母不为零，采取首先将文档数据 $\tilde{x}_i$ 投影到 PCA 子空间来消除平凡解。设 $\boldsymbol{W}_{\mathrm{PCA}}$ 表示 PCA 的变换矩阵，则经过 PCA 投影后，文档向量 $\tilde{\boldsymbol{x}}_i$ 变为 $\boldsymbol{x}_i$：

$$\tilde{\boldsymbol{x}}_i\rightarrow\boldsymbol{x}_i=\boldsymbol{W}_{\mathrm{PCA}}^{\mathrm{T}}\tilde{\boldsymbol{x}}_i \tag{17}$$

于是，原始文档向量 $\tilde{\boldsymbol{X}}=\left[\tilde{\boldsymbol{x}}_1,\tilde{\boldsymbol{x}}_2,\cdots,\tilde{\boldsymbol{x}}_n\right]$ 经 PCA 投影后变成 $\boldsymbol{X}=\left\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n\right\}$。

Step2 构造文档类内紧凑图和类间分离图。设 G 和 G^{p} 分别表示 2 个具有 n 个顶点的无向图：类内紧凑图和类间分离图，其中，顶点 i 对应于文档数据 x_i。

在文档类内紧凑图 G 中，对于任意文档数据 x_i 和 x_j，如果它们具有相同的类别，且满足 x_j 是 x_i 的 k_1－最近邻居，则在文档 x_i 和 x_j 之间创建一条边。

在文档类间分离图 G^{p} 中，对于任意文档数据元组 $\left(x_i,x_j\right)$，如果同时具备如下 2 个条件：(1) x_i 是 x_j 的 k_2－最近邻居，且 $x_i\left(\in C_1\right)$ 和 $x_j\left(\in C_2\right)$ 来自不同的类，即类 $C_1\neq C_2$；

(2) x_i 和 x_j 之间的距离是所有分别来自类 C_1 和 C_2 中的文档数据组成的 k_2 – 最近邻居元组中距离最短的。则在 x_i 和 x_j 之间创建一条边。

Step3 确定图 G 和 G^p 中边的权重。在类内紧凑图 G 中，如果顶点 i 和 j 之间有条边，则设其边权重 $S_{ij}=1$，否则为 0。在类间分离图 G^p 中，如果顶点 i 和 j 之间有条边，则设其边权重 $S_{ij}^p=1$，否则为零。

Step4 根据 MFA 的优化目标，确定高维文档数据的低维嵌入。首先依据式(2)计算出对角矩阵 $\boldsymbol{D}$ 和 $\boldsymbol{D}^p$，然后根据式(1)获得优化的 MFA 变换向量 $\boldsymbol{W}_{\mathrm{MFA}}=[\boldsymbol{w}_1,\boldsymbol{w}_2,\cdots,\boldsymbol{w}_d]$，其中，$\boldsymbol{w}_1,\boldsymbol{w}_2,\cdots,\boldsymbol{w}_d$ 是矩阵 $\left(\boldsymbol{X}\left(\boldsymbol{D}^p-\boldsymbol{S}^p\right)\boldsymbol{X}^{\mathrm{T}}\right)^{-1}\left(\boldsymbol{X}\left(\boldsymbol{D}-\boldsymbol{S}\right)\boldsymbol{X}^{\mathrm{T}}\right)$ 中对应于前 d 个最小特征值对应的特征向量。于是，高维文档数据的低维嵌入可描述为如下公式：

$$\tilde{\boldsymbol{x}}_i \to \boldsymbol{x}_i=\left(\boldsymbol{W}_{\mathrm{PCA}}\boldsymbol{W}_{\mathrm{MFA}}\right)^{\mathrm{T}}\tilde{\boldsymbol{x}}_i \tag{18}$$

其中，$\boldsymbol{x}_i$ 是 D -维文档数据 $\tilde{\boldsymbol{x}}_i$ 的 d -维表示，这里 $d \square D$。

Step5 利用经 MFA 降维后获得的 d -维文档数据 $\{x_i\}_{i=1}^n$ 依据式(4)、式(5)建立支持向量聚类算法 SVC 的优化目标函数。

Step6 SVC 的训练过程。依据式(6)建立 SVC 优化目标的拉格朗日函数，并通过式(7)~式(11)的计算过程，获得式(6)的 Wolfe 对偶形式(即式(12))，然后在式(13)的约束条件下获得支持向量，即当 $0<\beta_i<C$ 时，其所对应的数据点。

Step7 文档聚类过程。首先根据式(15)计算出支持向量 x_i 到球心的距离，即球半径：$R=R(x_i)$；其次，根据式(16)建立邻接矩阵 $\boldsymbol{A}$；最后，根据邻接矩阵 $\boldsymbol{A}$ 确定各个连通分量，连通分量的数目就是聚类的数目，一个连通分量对应于一个聚类。对于那些包含一个数据点的连通分量，可标记为噪声或者孤立点。

4 实验结果

为了测试本文提出的基于间隔流形学习 MFA 和支持向量聚类 SVC 的文档聚类算法(简称 MFA-SVC)性能，将其与常用的文档聚类算法：潜在语义索引法(LSI)，拉普拉斯特征映射法(LE)[1]，非负矩阵分解法(NMF)[2]及局部保持索引法(LPI)[3]进行了比较。MFA-SVC 算法中的参数设置为：类内最近邻居数 k_1 和类间最近邻居数 k_2 分别设置为 6 和 10，核函数 $K(.)$ 采用常用的高斯核函数 $K(x,y)=e^{-\gamma\|x-y\|^2}$，模型参数采用 10 次交叉验证法来确定。

测试数据采用文档聚类研究中常用的测试数据集 Reuters-21578 和 TDT2。文档聚类算法的性能评价指标采用了常用的文档聚类准确率和归一化互信息量[2]，这 2 个评价指标的值越大，说明聚类算法的性能越好。计算方法如下：

给定文档 x_i，设其类别 r_i 和 s_i 分别代表由文档聚类算法获得类别及测试数据集本身给定的类别，则文档聚类准确率的计算公式如下：

$$\text{聚类准确率}=\frac{\sum_{i=1}^{n}\delta\left(s_i,map\left(r_i\right)\right)}{n} \tag{19}$$

其中，n 表示所有的文档数；$\delta\left(s_i,map\left(r_i\right)\right)$ 的定义如下：

$$\delta\left(s_i,map\left(r_i\right)\right)=\begin{cases}1 & s_i=map\left(r_i\right)\\ 0 & \text{otherwise}\end{cases} \tag{20}$$

其中，$map\left(r_i\right)$ 表示把类别 r_i 映射到数据集中的等价类别所需要的置换映射函数。

设 C 和 C' 分别表示由基本的 K-均值聚类算法及其他文档聚类算法得到的聚类集合，则它们之间的互信息量 $MI(C,C')$ 定义如下：

$$MI(C,C')=\sum_{c_i\in C,c_j'\in C'}p\left(c_i,c_j'\right)\mathrm{lb}\frac{p\left(c_i,c_j'\right)}{p\left(c_i\right)p\left(c_j'\right)} \tag{21}$$

其中，$p\left(c_i\right)$ 和 $p\left(c_j'\right)$ 表示任一文档分别属于聚类 c_i 和 c_j' 的概率，$p\left(c_i,c_j'\right)$ 表示任一文档同时属于聚类 c_i 和 c_j' 的联合概率。于是，归一化的互信息量 $NMI(C,C')$ 定义如下：

$$NMI(C,C')=\frac{MI(C,C')}{\max\left(H(C),H(C')\right)} \tag{22}$$

其中，$H(C)$ 和 $H(C')$ 分别表示聚类 C 和 C' 的熵。

上述 5 种文档聚类算法在测试文档数据集 Reuters-21578 和 TDT2 上的聚类性能比较如表 1、表 2 所示。

表 1 在 Reuters-21578 上的聚类性能比较

聚类算法	聚类准确率	归一化互信息
LSI	0.692	0.574
LE	0.732	0.618
NMF	0.729	0.607
LPI	0.773	0.650
MFA-SVC	0.796	0.684

表 2 在 TDT2 上的聚类性能比较

聚类算法	聚类准确率	归一化互信息
LSI	0.883	0.892
LE	0.990	0.971
NMF	0.898	0.859
LPI	0.992	0.973
MFA-SVC	0.994	0.976

从表 1 和表 2 中的实验结果可以得出如下结论：

(1)本文提出的 MFA-SVC 算法在 Reuters-21578 和 TDT2 上的聚类准确率和归一化互信息量均优于常用的 LSI、LE、NMF 和 LPI。原因在于：在维数降维的过程中 MFA 算法能够同时利用类内紧凑图和类间分离图来显式地描述类内文档数据的相似性及不同类间文档数据的分离性；另外，SVC 聚类算法无需预先指定聚类的数目，且不受聚类形状的限制，因此能够适应处理复杂文档数据的需要。所以，MFA-SVC 算法取得了很好的聚类效果。

(2)LSI 和 NMF 算法的聚类性能较差，原因在于：LSI 和 NMF 算法旨在以最小化全局构造误差为目标，用于发现近似于原始高维文档空间的子空间表示，因此，对于以区分不同类别的文档为主要目的的文档聚类任务而言是非优化的。

(3)LE 和 LPI 算法虽然能够在维数降维的过程中保持局部流形结构，但是 LE 算法没有显式地给出新测试数据的低维表示，而需要利用全部数据来获得新测试数据的低维表示。LPI 算法的不足之处在于：它只是采用单一的邻接图来实现类内文档数据之间的相似性，而没有显式地考虑不同类间文档数据的分离性。因此，对于文档聚类而言也是非优化的。

5 结束语

针对一般聚类算法在处理高维文档数据时面临的“维数灾难”和聚类准确率不高的问题，本文提出一种基于间隔流形学习 MFA 和 SVC 的文档聚类算法。实验结果表明，该算法的聚类性能优于其他常用的基于文档聚类算法。

（下转第 48 页）