一种半监督局部线性嵌入算法的文本分类方法 *

夏士雄, 李佑文, 周 勇
(中国矿业大学 计算机科学与技术学院, 江苏 徐州 221116)

摘 要: 针对局部线性嵌入算法(LLE)应用于非监督机器学习中的缺陷,将该算法与半监督思想相结合,提出了一种基于半监督局部线性嵌入算法的文本分类方法。通过使用文本数据的流形结构和少量的标签样本,将LLE中的距离矩阵采用分段形式进行调整;使用调整后的矩阵进行线性重建从而实现数据降维;针对半监督LLE中使用欧氏距离的缺点,采用高斯核函数将欧氏距离进行变换,并用新的核距离取代欧氏距离,提出了基于核的半监督局部线性嵌入算法;最后通过仿真实验验证了改进算法的有效性。

关键词: 局部线性嵌入算法; 半监督学习; 流形学习; 文本分类; 核函数
中图分类号: TP311 **文献标志码**: A **文章编号**: 1001-3695(2010)01-0064-04
doi:10.3969/j.issn.1001-3695.2010.01.018

Method based on semi-supervised local linear embedding algorithm for text classification

XIA Shi-xiong, LI You-wen, ZHOU Yong
(*School of Computer Science & Technology, China University of Mining & Technology, Xuzhou Jiangsu* 221116, *China*)

Abstract: In order to solve the defects of local linear embedding algorithm(LLE) could only be used in unsupervised machine learning, combined this algorithm and the thinking of semi-supervised learning together, this paper proposed a method based on semi-supervised local linear embedding algorithm for text classification. Firstly, with the manifold structure of text data and some labeled samples, this algorithm revised the distance matrix in LLE algorithm by using piecewise function. Secondly, in order to achieve the purpose of dimensionality reduction, reconstructed the samples linearly by using the adjusted matrix. Then, because of shortcomings of the Euclidean distance in semi-supervised local linear embedding algorithm, improved it by proposing kernel based semi-supervised local linear embedding algorithm, which transformed and replaced Euclidean distance by Gaussian kernel function distance. Finally, the results of simulated experiments indicate these algorithms can really promote the performance of text classification.

Key words: LLE; semi-supervised learning; manifold learning; text classification; kernel function

0 引言

随着当今网络社会的飞速发展,人们获取的数据信息急剧增加,而且这些数据包含的信息包罗万千,尤其是以文字形式出现的信息更加丰富。然而如何有效地从这些文本信息中获取有效的信息变得尤为重要,因此,文本分类问题是数据挖掘中的一个重要研究内容。在数据挖掘中,分类属于监督学习,而绝大多数的有监督学习方法都依赖于训练样本集,却忽略了未标注样本的重要作用,由于获取标记数据样本的代价过高,如何有效地利用少量标注样本进行训练就变得尤为重要。将这种利用少量标记样本协助训练的算法称之为半监督学习算法[1](semi-supervised learning algorithm)。

在半监督学习中有两个基本的假设,即聚类假设(cluster assumption)和流形假设[2](manifold assumption)。聚类假设是指在同一聚类簇内的样本有较大的可能性属于同一个类别。未标记样本可以帮助探明样本空间中数据分布稀疏和稠密的区域,从而指导学习算法,使其尽量通过数据稀疏的区域。流形假设是指处于一个很小的局部邻域内的样本具有相似的性质,因此其标记也应该相似。该假设反映了决策函数的局部平滑特性。在该假设下,由于加入了大量的未标记样本,使得样本空间变得更加稠密,从而可以更加准确地描述局部区域的特性,使得决策函数能够更好地进行数据拟合。

在进行文本分类时,文本向量的维数很高。随着维数的增加,算法的性能将变得很差,因此文本向量的维数约简变得相当重要。针对这种非线性结构的数据降维,一般都是使用流形学习算法,这正好与本文使用半监督学习进行文本分类的流形假设一致。因此本文也采用流形学习的降维方法对文本向量进行降维,并通过实验结果证明了该方法的可行性。

目前,比较有代表性的流形学习方法有等距离映射算法[3]、局部线性嵌入算法[4,5]和拉普拉斯本征映射[6]等。其中,尤其以局部线性嵌入算法应用较多,但是该算法的很多应用都局限在进行人脸识别时的图像向量的降维[4~6],很少使用在文本向量的维数约简中。针对这一现状,在比较图像向量和文本向量的基础上,基于文本向量和图像向量都是高维非线性

收稿日期: 2009-04-23; **修回日期**: 2009-06-26 **基金项目**: 国家自然科学基金资助项目(50674086);高等学校博士学科点专项科研基金资助项目(20060290508)

作者简介:夏士雄(1961-),男,黑龙江鹤岗人,教授,博导,主要研究方向为模式识别、人工智能等(xiasx@cumt.edu.cn);李佑文(1985-),男,湖北监利人,硕士研究生,主要研究方向为数据挖掘、计算机应用技术等;周勇(1974-),男,江苏徐州人,副教授,博士,主要研究方向为人工智能、无线传感器网络.

向量的共同点,本文提出了一种将 LLE(local linear embedding,局部线性嵌入)算法用于文本向量降维的方法。然而经典的 LLE 算法是一种无监督的流形学习算法,本文中使用的却是半监督学习样本,只有部分样本带有标记,因此必须对 LLE 算法进行改进,使之能够适应半监督流形学习。另外,由于文本向量一般都分布在高维稀疏的空间中,为了能在降维后尽可能保持原有样本数据的拓扑结构,本文还使用一种核的方法来计算 LLE 算法中的距离。

1　文本分类的预处理

在几乎所有的数据挖掘算法中,样本的预处理都是十分重要的,预处理所花费的代价可能高达 70%,所以在进行文本分类时,关于文本的预处理也相当重要。

对于文本分类,首先要解决的问题是如何在计算机中表示文本,使计算机能够识别文本。常用的文本表示模型有布尔逻辑模型、概率模型和向量空间模型(vector space model,VSM)等[7]。在进行文本分类时,VSM 在表示方法上有巨大的优势,其基本思想是将文本表示成一个向量,成为多维空间中的一个点,然后再使用经典的分类算法将其进行分类,本文也使用向量空间模型表示文本。

如何将一篇文档表示成为一个合适的向量,需要使用到很多技术,具体包括分词、去停用词、特征抽取、权值计算。分词一般是指将没有明显分界标志的文本内容分割成机器词典中存在的词语,在进行分词时可以对没有具体意义的停用词进行过滤以去除那些对分类没有实际影响的词语。特征抽取是指在不影响分类效果的情况下,从原来的特征集合 T 中抽取出一个子集 T',从而降低维数。本文使用改进的局部线性嵌入算法来进行降维。在降维后得到的都是对分类有较高帮助的特征词汇,最后为这些特征词汇计算对应的权值,本文采用信息增益方法计算权值[8],将最后计算得到的权值和对应的特征词保存在 XML 文件中。到此文本分类的预处理过程就完成了,后续的文本分类算法可以直接使用 XML 文件中保存的文本向量进行分类处理。

2　局部线性嵌入算法

局部线性嵌入算法是由 Saul 等人[4,5]提出的一种新的专门针对非线性数据降维的方法,通过 LLE 算法处理后的低维数据能够较好地保持原有数据在高维空间中的拓扑结构。LLE 算法是一种功能很强大的非线性数据降维方法。它基于流形假设的思想,利用局部的线性来逼近全局的非线性,通过保持局部的几何结构不变以及局部的相互叠加来提供整体的信息,从而保持整体的集合性质,即 LLE 算法认为在流形上每一个局部邻域内的任意一点都可以由邻域内的其他点线性叠加表示。它基于高维空间中相邻或相关的点映射到低维空间之后应该也是相邻或相关的思想,将高维空间中的点映射到低维空间中取得了相当大的成功。经过几年的发展,现在 LLE 算法已经广泛应用于图像数据的分类与聚类、文字识别、手写体识别、多维数据的可视化以及生物信息学等领域中。

LLE 算法的目的是将高维空间 R^m 中的一个样本数据点 $X=\{x_1,x_2,\cdots,x_m\}$,嵌入到一个低维空间 $R^d(d<<m)$ 中,得到 X 降维后对应的点集 $Y=\{y_1,y_2,\cdots,y_d\}$,在映射过程中尽量使损失达到最小。LLE 算法的具体步骤(图 1)主要有三步[4,5,9]:

a)选择高维空间中每个样本点 $x_i(i=1,\cdots,n)$ 的 K 个近邻样本。常常采用的是计算每个点与其他点的欧氏距离 $d_{ij}=\sqrt{\sum_{k=1}^{m}(X_{ik}-X_{jk})^2}$,形成距离矩阵,从矩阵中为每个点选择距离最近的 K 个点组成集合记为 $\mathrm{KNN}(x_i)$。其中 K 是预先指定的参数。

b)对于每个样本点 $x_i(i=1,\cdots,n)$,寻找一个 K 维的权重向量 $W_{ij}(j=1,\cdots,K)$(满足 $\sum_{j=1}^{K}W_{ij}=1$)使得 x_i 在由其 K 个近邻点 $\mathrm{KNN}(x_i)$ 线性表示时的损失达到最小,即

$$\varepsilon(W)=\min\Big(\sum_{i=1}^{n}\Big|x_i-\sum_{j=1,x_j\in\mathrm{KNN}(x_i)}^{K}W_{ij}x_j\Big|^2\Big)\tag{1}$$

c)将所有高维样本 $x_i\in R^m$ 映射到低维空间中 $y_i\in R^d$ $(d<<m)$,映射权重向量不变,使在低维空间中的损失最小即可,即

$$\varepsilon(Y)=\min\Big(\sum_{i=1}^{n}\Big|y_i-\sum_{j=1}^{K}W_{ij}y_j\Big|^2\Big)\tag{2}$$

其中:$\sum_{i=1}^{n}y_i=0$,$\frac{1}{N}\sum_{i=1}^{n}y_iy_i^{\mathrm{T}}=I$,$I$ 为 $d\times d$ 的单位矩阵。

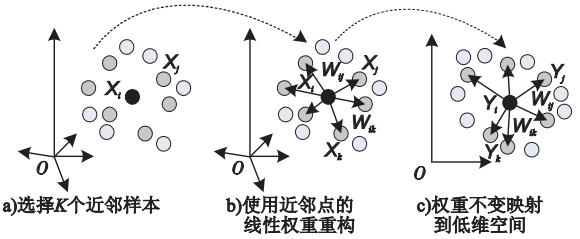

图1　LLE算法的步骤

在求解 LLE 算法的步骤 c)时,可以将 $n\times K$ 的权重向量矩阵 $W_{n\times K}$ 扩充为 $n\times n$ 的稀疏矩阵 $W_{n\times n}$,只需使得 x_j 是 x_i 的近邻点时,W_{ij} 的值不变;当 x_j 不是 x_i 的近邻点时,$W_{ij}=0$ 即可。若此时定义一个新的对称矩阵 $M=(I-W_{n\times n})^{\mathrm{T}}(I-W_{n\times n})$,则式(2)可以改写为

$$\varepsilon(Y)=\min\Big(\sum_{i=1}^{n}\sum_{j=1}^{n}M_{ij}y_i^{\mathrm{T}}y_j\Big)\tag{3}$$

通过求解知道,要使式(3)获得最小值,只需取 Y 为 M 的最小 d 个非零特征值所对应的特征向量即可。由于 M 的最小特征值一般近似于 0,通常在对特征值从小到大排序后取 $2\sim d+1$ 个特征值所对应的特征向量。

3　基于半监督局部线性嵌入算法的文本分类方法

3.1　监督局部线性嵌入算法(supervised LLE,SLLE)

LLE 算法本身是一种无监督的降维算法,但是随着 LLE 算法的逐渐发展,Ridder 等人[10]提出一种有监督的 LLE 算法。有监督学习算法中的带标签训练数据可以指导传统的 LLE 算法去寻找每个样本点的 K 个邻近点,所以文献[10]在计算点与点之间的距离时采用了一种距离矩阵调整策略:

$$D'=D+\alpha\max(D)\Delta\tag{4}$$

其中:D 是点与点间的欧氏距离矩阵,$\max(D)$ 表示的是类与类之间的最大欧氏距离;在计算新的距离矩阵 D' 时,若两点属于同一类,则 $\Delta=0$,否则 $\Delta=1$;$\alpha(0<\alpha<1)$ 是一个经验参数,用于控制空间内点的稀疏稠密距离参数,当 $\alpha=0$ 时,SLLE 算法与 LLE 算法相同。SLLE 算法的其他两个步骤与 LLE 算法的完全相同,Ridder 他们的实验证明了 SLLE 算法的可行性。

3.2　半监督局部线性嵌入算法(semi-supervised LLE,SSLLE)

针对目前很少有将局部线性嵌入算法引入到半监督学习

中的现状,并且由于在使用半监督算法时就已经存在了流形假设的前提,本文将半监督与 LLE 算法相结合,提出了一种新的基于半监督的局部线性嵌入算法 SSLLE,该算法使用部分带标签的样本来指导 LLE 算法。在 SLLE 算法中所有的样本都是带有标签的,所以在计算新的距离矩阵时很容易进行调整。然而由于 SSLLE 算法中只有部分样本是带标签的,不能计算类与类之间的最大欧氏距离 max(D),不能像 SLLE 算法一样,可以采用下面的策略来调整距离矩阵从而指导 LLE 算法准确地寻找 K 个近邻点。

$$D'=\begin{cases} d_{ij}-\gamma e^{-d_{ij}} & ①x_i,x_j \text{ 都带标记,并且属于同一类} \\ d_{ij}+\gamma e^{-d_{ij}} & ②x_i,x_j \text{ 都带标记,并且不属于同一类} \\ d_{ij}+\gamma/2\ e^{-d_{ij}} & ③x_i,x_j \text{ 仅一个带标记,但 } x_i\in \mathrm{KNN}(k_j) \text{ 或 } x_j\in \mathrm{KNN}(x_i) \\ d_{ij} & ④\text{其他} \end{cases} \quad (5)$$

其中:$\gamma(0<\gamma<1)$ 是调节系数,用于调节 SSLLE 算法中带标签样本和不带标签样本在被选择为近邻点时的可能性。

对于式(5),本文给出如下的定性分析:在算法第一次选择 K 个近邻点时按照 LLE 算法的步骤 a)进行,然后再使用式(5)调整距离矩阵重新选择近邻点。为了加强半监督算法中带标签样本的指导意义,在选择近邻点时,应该偏向于选择带标签的样本以增加算法的准确性。在标记样本选择 K 个近邻时,其近邻点可能是标记的样本,该标记的近邻点可能与原样本属于同一类,对于这样的近邻点应该有更大的可能性被选中,即式(5)中的①;该标记的近邻点也可能与原样本不属于同一类,对于这样的近邻点被选中的可能性应该较小,即式(5)中的②;若在标记样本选择 K 个近邻时,其近邻点是未标记的样本,但在第一次选择时将其选为了近邻点,对于这样的点在最后是否被选择为近邻点,采取式(5)中的③较好;对于其他的情况(包括 x_i,x_j 都不带标记和③中仅一个带标记,但另一个也不是其 K 近邻点的情况),采取式(5)中的④,其距离不予调整。

选择了每个样本的 K 个近邻后,SSLLE 算法的其他两个步骤与 LLE 算法相同,最后实验证明该算法可以在一定程度上改善分类性能。

3.3　核半监督局部线性嵌入算法(kernel based semi-supervised LLE,K-SSLLE)

以上提到的三个算法,无论是 LLE、SLLE 还是 SSLLE 算法,都采用的是欧氏距离来计算两个样本之间的距离。然而对于具有相同欧氏距离长度的样本所在的位置信息却忽略了,这是欧氏距离的最大缺点之一。另外,由于文本向量一般都是分布在高维稀疏的空间中,考虑到使用核函数可以尽可能地保持高维空间中样本点的拓扑信息,本文将核函数的思想和半监督局部线性嵌入算法结合起来,提出了一种采用核距离函数的 SSLLE 算法,该算法能更好地保留空间样本在降维后的信息。同 SLLE 和 SSLLE 算法一样,K-SSLLE 算法也只是改进了 LLE 算法的第 a)步,即使用核距离代替欧氏距离来寻找 K 个近邻点,K-SSLLE 算法的后两个步骤与 SSLLE 算法一致。

目前,使用最广泛的核函数有三种[11],即多项式核函数、高斯核函数和 sigmoid 核函数。但是目前核函数的选择依据尚没有定论,在机器学习模型中,一般是凭经验选取。由于高斯核函数是比较普遍使用的核函数,本文也采用高斯核函数将欧氏距离转换成核距离,如式(6)所示:

$$d'_{ij}=\mathrm{dist}(\phi(x_i),\phi(x_j))=\sqrt{\|\phi(x_i)-\phi(x_j)\|^2}=\sqrt{K_{ii}-2K_{ij}+K_{jj}} \quad (6)$$

其中:K 即核函数,取 $K(x,y)=\exp(-\beta\|x-y\|^2)$ 时为高斯核函数。K-SSLLE 算法将式(5)中的欧氏 d_{ij} 距离换成 d'_{ij},其他步骤与 SSLLE 算法相同。最后实验结果表明,将高斯核函数引入到 SSLLE 算法后要比单纯的 SSLLE 算法性能好。

3.4　核半监督局部线性嵌入文本分类算法的步骤

由于 K-SSLLE 算法只是改进了 SSLLE 算法距离计算,其他两个步骤都相同,在此仅给出核半监督局部线性嵌入文本分类算法的步骤描述。

输入:C_i 类标记文档$\{d_{i1},d_{i2},\cdots,d_{iNei}\}$一共 N 篇以及大约 $8N$ 的未标记文档,测试文档集 $S=\{S_1,S_2,\cdots,S_n\}$。

输出:S_i 所属的类别 C_i。

a)对所有输入文档进行分词、去停用词等预处理,将得到的特征词和词频保存在 XML 文件中。

b)提取每个 XML 文件中的原始文本向量,使用欧氏距离公式计算所有文本之间的距离,找出每个样本点的初始 K 邻近样本集合。

c)使用核函数(式(6))重新计算所有文本之间的距离,并利用初始 K 邻近样本集合和式(5)调整距离矩阵,然后重新为每个样本点选择 K 邻近样本集合。

d)使用式(1)为每个样本点寻找权重向量,使其在由其 K 邻近样本线性表示时的损失最小。

e)使用式(2),利用步骤 d)中的权重向量在低维空间中线性重建原高维向量,并满足在低维空间中的损失最小从而达到降维的目的,将降维后的特征词和词频保存在 XML 文件中。

f)使用 EM 半监督算法,利用步骤 e)中保存的文本特征向量将测试文档 S_i 依次分类,并输出其类别属性 C_i。

4　实验与分析

4.1　实验数据和环境

本文使用的实验数据来自复旦大学计算机信息与技术系国际数据库中心自然语言处理小组[12]。其中 answer. rar 为测试语料,共 9 833 篇文档;train. rar 为训练语料,共 9 804 篇文档,分为 20 个类别。然而该语料库过大,如果采用全部的文档进行实验,将会影响分类效率。为验证本文算法的有效性,只选取其中的部分语料,具体如表 1 所示。

表 1　实验数据

类别	训练集	测试集	类别	训练集	测试集
C1-art	240	30	C5-economy	220	27
C2-space	240	30	C6-politics	240	30
C3-computer	250	31	C7-sports	260	33
C4-enviornment	240	30	总计	1 690	211

实验平台:CPU 为 Intel Pentium Dual Core Processor,规格 Genuine Intel® CPU 2140 1.6 GHz,内存为 2 GB DDR2;实验开发环境:Windows XP + Microsoft Visual Studio 2005 C#. NET。

4.2　实验与分析

本文的主要工作集中在文本向量的降维上,具体使用哪种文本分类器不再具体讨论。另外,本文在评价文本分类指标时综合考虑分类的准确率 precision 和召回率 recall[13,14],并将准确率与召回率结合在一起得到综合指标 F_1 值,$F_1=(2\times \mathrm{precision}\times \mathrm{recall})/(\mathrm{precision}+\mathrm{recall})$。针对以上提到的几个算法,设计如下四个对比实验(其中参数 K,α,γ 都经过多次测试,取结果较好时的值)。

实验 1　不进行降维时进行文本分类(数据集是全标记的,即表 1 中的训练数据全都带标记),使用 KNN 分类器,实验

结果如表 2 所示。

实验 2　使用 SLLE 算法降维后进行文本分类(数据集同实验 1),使用 KNN 分类器,实验结果如表 3 所示。

表 2　实验 1 结果

类别	precision	recall	F_1 值
C1-art	0.832	0.845	0.838
C2-space	0.886	0.871	0.878
C3-computer	0.825	0.833	0.829
C4-enviornment	0.849	0.842	0.845
C5-economy	0.902	0.898	0.90
C6-politics	0.875	0.846	0.860
C7-sports	0.906	0.86	0.882
平均值	0.868	0.856	0.862

表 3　实验 2 结果

类别	precision	recall	F_1 值
C1-art	0.844	0.832	0.838
C2-space	0.86	0.884	0.872
C3-computer	0.853	0.842	0.847
C4-enviornment	0.844	0.823	0.833
C5-economy	0.885	0.901	0.8931
C6-politics	0.848	0.862	0.855
C7-sports	0.896	0.858	0.877
平均值	0.861	0.857	0.859

对于实验 1 没有进行降维处理,在进行实验时,从程序里面了解到文本向量的维度大约是 12 000 维;而实验 2 中,取参数 $K=30$,$\alpha=0.01$ 时,文本向量的维度大约是 400 维左右,实验 2 通过 SLLE 降维处理后在很大程度上加快了算法的运行。通过分析以上两个实验结果,发现 KNN 文本分类的算法的正确率都是 86% 左右,在准确率和召回率以及 F_1 值上都没有很大的波动,这说明使用 SLLE 算法进行降维后基本没有影响算法的性能,从而为后面的实验使用改进 LLE 算法提供了可行性证明。

实验 3　使用 SSLLE 算法降维后进行文本分类(数据集是部分标记的,将表 1 中的每类训练数据随机标记 1/8,其余 7/8 取消其标记)。由于训练数据是半监督数据,可以使用一种常见的半监督算法进行分类——EM 算法[15]。实验结果如表 4 所示。

实验 4　使用 K-SSLLE 算法降维后进行文本分类(数据集同实验 3),使用 EM 算法进行分类,实验结果如表 5 所示。

表 4　实验 3 结果

类别	precision	recall	F_1 值
C1-art	0.862	0.874	0.868
C2-space	0.876	0.877	0.876
C3-computer	0.937	0.912	0.924
C4-enviornment	0.838	0.84	0.839
C5-economy	0.856	0.895	0.875
C6-politics	0.874	0.898	0.886
C7-sports	0.921	0.888	0.904
平均值	0.881	0.883	0.882

表 5　实验 4 结果

类别	precision	recall	F_1 值
C1-art	0.891	0.894	0.892
C2-space	0.865	0.861	0.863
C3-computer	0.933	0.919	0.926
C4-enviornment	0.906	0.934	0.920
C5-economy	0.913	0.884	0.898
C6-politics	0.886	0.909	0.897
C7-sports	0.933	0.928	0.930
平均值	0.904	0.904	0.904

首先对比实验 2 和 3(不考虑其使用不同分类算法)。实验 3 中取参数 $K=28$,$\gamma=0.02$。实验 2 中采用全标记样本,而实验 3 只采用了实验 2 中约 1/8 的标记样本,这在很大程度上节约了获取标记样本的代价。比较这两个实验的结果发现,在使用 SSLLE 算法后,文本分类的准确性也比使用 SLLE 算法有大约 2.5% 的提高,这从两个方面说明将半监督思想和 LLE 算法结合在一起取得了成功。

再对比实验 3 和 4(实验 4 中参数取值同实验 3)。实验 4 比 3 的准确性提高了 1.2% 左右,这说明使用核函数的确可以在一定程度上改进高维空间中点的距离计算,即核函数可以更好地保留高维空间中样本点的拓扑结构信息,证明了本文算法中引入高斯核函数确实提高了算法的性能。

5　结束语

局部线性嵌入算法作为一种无监督的非线性数据降维方法,在很多机器学习领域都有很广泛的应用,本文率先将它引入到文本分类的领域。基于 LLE 算法,本文提出了一种基于半监督的局部线性嵌入算法 SSLLE,并随后又将 SSLLE 算法扩展成了基于核的半监督局部线性嵌入算法 K-SSLLE。这两种算法都针对半监督训练样本集,并充分利用半监督训练数据集中的部分带标签样本,通过 LLE 算法,使用一些已标记的样本可以很好地构造样本点的局部邻域特征向量,能够有效地将高维空间中的点嵌入到低维空间中,有利于对数据进行分类,最后的实验也证明了这一点。

在 SSLLE 算法和 K-SSLLE 算法中也存在一些需要改进的地方。其中,如何选择近邻参数 K 就是一个很值得研究的问题。如果 K 值过大,即选取了较多的近邻点,则这些大量的最近邻域可能会促成流形小规模结构的消除以及整个流形的平滑;相反,K 值太小,太少的邻域可能误将连续的流形结构划分成脱节的子流形。另外,关于调节参数 α、γ 的选取也都是值得研究的问题,这都是下一步工作需要研究的重点。

参考文献:

[1] ZHU Xiao-jin. Semi-supervised learning literature survey, TR 1530 [R]. Madison: University of Wisconsin-Madison, 2007.

[2] 周志华. 半监督学习中的协同训练风范[M]//周志华,王珏. 机器学习及其应用. 北京:清华大学出版社, 2007:259-275.

[3] BALASUBRAMANIAN M, SCHWARTZ E L. The Isomap algorithm and topological stability[J]. Science, 2002,295(5552):7-7.

[4] ROWEIS S T, SAUL L K. Nonlinear dimensionality reduction by locally linear embedding[J]. Science, 2000, 290(5500): 2323-2326.

[5] SAUL L K, ROWEIS S T. An introduction to locally linear embedding [EB/OL]. (2001). http://www.cs.toronto.edu/~roweis/lle/.

[6] TENENBAUM J B, De SILVA V, LANGFORD J C. A global geometric framework for nonlinear dimensionality reduction[J]. Science, 2000,290(5500):2319-2323.

[7] 苏金树,张博锋,徐昕. 基于机器学习的文本分类技术研究进展[J]. 软件学报,2006,17(9):1848-1859.

[8] 陆玉昌,鲁明羽,李凡. 向量空间法中单词权重函数的分析和构造[J]. 计算机研究与发展,2002,39(10):1205-1210.

[9] 王和勇,郑杰,姚正安,等. 基于聚类和改进距离的 LLE 方法在数据降维中的应用[J]. 计算机研究与发展,2006,43(8):1485-1490.

[10] De RIDDER D, KOUROPTEVA O, OKUN O, *et al*. Supervised locally linear embedding[M]//Artificial Neural Networks and Neural Information Processing ICANN/ICONIP. Berlin: Springer, 2003:333-341.

[11] 张莉,周伟达,焦李成. 核聚类算法[J]. 计算机学报,2002,25(6):587-590.

[12] 中文自然语言处理开放平台[EB/OL]. [2007-08-30]. http://www.nlp.org.cn/docs/doclist.php? cat_id=16.

[13] YANG Yi-ming, PEDERSEN J O. A comparative study on feature selection in text categorization[C]//Proc of the 14th International Conference on Machine Learning. San Francisco: Morgan Kaufmann Publishers, 1997:412-420.

[14] WANG Xin-hao, LUO Ding-sheng, WU Xi-hong, *et al*. Improving Chinese text categorization by outlier learning[C]//Proc of IEEE International Conference on Natural Language Processing and Knowledge Engineering. 2005:602-607.

[15] NIGAM K, McCALLUM A K, THRUN S, *et al*. Text classification from labeled and unlabeled documents using EM[J]. Machine Learning, 2000,39(2):103-134.