文章编号:1671-6833(2012)05-0121-04

基于流形学习的基因微阵列数据分类方法

李 强,石陆魁,刘恩海,王 歌

(河北工业大学 计算机科学与软件学院,天津 300401)

摘 要:提出了一种结合流形学习方法与分类算法的基因微阵列数据分类模型,先用流形学习算法对基因微阵列数据进行降维处理,然后再对降维后的数据进行分类.在实验中将流形学习算法 LLE、ISOMAP、LE 和 LTSA 与三种分类算法相结合,并与直接用高维数据进行分类的结果进行了比较,实验结果表明所提出的模型极大地提高了分类精度,同时也提高了分类算法的执行效率.

关键词:流形学习;分类;基因;微阵列数据

中图分类号:TP181 **文献标志码**:A doi:10.3969/j.issn.1671-6833.2012.05.027

0 引言

近些年来,随着基因微阵列数据应用逐渐趋于广泛和微阵列数据库的不断完善,人们越来越需要充分深入地从大量的数据中捕获信息.即使再大的基因组,人们可能只获得少部分的基因信息.如果能研究出更加先进的基因分析工具,使人们从基因微阵列数据中提取出有用的甚至于更深层次的信息,如基因功能信息、基因微阵列的进化信息、与疾病相关的信息等,无疑能发挥出至关重要的作用.因此,如何对这些复杂的数据进行有效地分析,挖掘出其中蕴含的有用信息成为当今社会研究的重点课题之一[1-7].

对基因微阵列数据进行分类是挖掘微阵列数据中有用信息的一种重要手段.目前,在基因微阵列数据分类研究方面,多数研究者采用有监督的分类方法,主要包括 K-近邻(K-Nearest Neighbor, K-NN)、支持向量机(Support Vector Machine, SVM)和朴素贝叶斯(Naive Bayes, NB)等方法[8].然而基因微阵列数据具有样本少、非线性、维数高等特点,一般每个样本的维数都高达几千甚至上万,这种特性导致目前的一些分类算法对其进行分类的效果不尽如人意.因此,如果能对微阵列数据集先进行特征选择或提取(降维处理),提取出与样本类别相关的基因或信息,再将分类方法应用到降维后的数据,可能会取得好的识别效果.而流形学习算法作为一种非线性降维方法,可以发现隐藏在高维数据中的非线性流形,在模式识别、数据挖掘等领域得到了广泛地应用.为此,作者提出了一种基于流形学习的基因微阵列数据分类方法,可以较大地提高分类精度.

1 相关工作

1.1 微阵列数据分类方法

在基因微阵列数据的分类研究中,多数研究人员采用有监督的分类方法,常用 K-NN、朴素贝叶斯和 SVM 等方法对微阵列数据进行分类.假设 X 是一个输入向量,K-NN 算法首先在训练集中找出与其距离最近的 K 个点,K 一般为奇数(避免二义性问题),然后再根据近邻点所属类别确定其类别.如果这 K 个点中属于某个类的点最多,则 X 就属于该类.可见 K-NN 算法的关键就是求样本点与训练集中每个点之间的距离,可以选择欧式距离、向量夹角余弦、Pearson 相关系数和 Minkowski 距离等.

贝叶斯分类算法利用概率统计知识进行分类.假定有 m 个类,分别用 $C_1, C_2, \cdots, C_m$ 表示,设 $X=\{x_1, x_2, \cdots, x_n\}$ 是一个未知的数据样本,X 属于类 C_i 当且仅当 $P(C_i|X) > P(C_j|X)$,其中 $1 \leqslant j \leqslant m$ 且 $j \neq i$.根据贝叶斯定理,$P(X)$ 对于所有类都为常数,最大化后验概率 $P(C_i|X)$ 可转化为最大化先验概率 $P(X|C_i)P(C_i)$.根据贝叶斯方

收稿日期:2011-02-20;**修订日期**:2011-05-20

基金项目:天津市应用基础及前沿技术研究计划资助项目(10JCZDJC16000)

通信作者:石陆魁(1974-),男,河北省邯郸人,河北工业大学副教授,博士,主要从事机器学习、数据挖掘研究.E-mail:shilukui@scse.hebut.edu.cn.

法,对一个未知类别的样本 X,可以先分别计算出 X 属于每个类别 C_i 的概率 $P(X|C_i)P(C_i)$,然后选择其中概率最大的类别作为其类别.

支持向量机遵循结构化风险最小化原则来解决分类问题.支持向量机的基本思想是:首先将输入空间投影到一个高维空间,然后在高维空间中基于分类间隔最大求得最优线性分类面.但由于支持向量机算法通过变换空间的维数不能反映出所获得分类器的复杂度,该算法所获得分类器的复杂度通过采用支持向量的个数来反映,这就避免了其它算法可能会产生的过拟合问题.

1.2 流形学习算法

流形学习可以发现隐藏在高维数据中的非线性流形,近年来得到了快速发展,并被应用到图像处理、模式识别等领域.比较具有代表性的流形学习算法包括局部线性嵌入法(Locally Linear Embedding,LLE)[9]、等度规映射法(ISOmetric feature MAPping,ISOMAP)[10]、拉普拉斯特征映射法(Laplacian Eigenmaps,LE)[11]和局部切空间校正法(Local Tangent Space Alignment,LTSA)[12]等.

LLE 算法将全局非线性转化为局部线性,其基本假设是每个数据点和它的邻域点位于流形的一个线性或几乎线性区域中,这样可以在数据集中的每一个样本点和它的邻域点之间构造局部线性平面,进而在此基础上建立函数并且优化[9].

ISOMAP 算法是对多维尺度分析(MultiDimensional Scaling,MDS)法的一种扩展,其基本思想是用测地线距离代替 MDS 中的欧式距离.算法的关键是计算所有点间的测地线距离,对于近邻点直接用欧式距离近似测地线距离,对于非近邻点用两点之间最短路径来近似测地线距离.算法包括三个步骤:第一,确定每个样本的 k 个近邻点,构建邻域图;第二,在邻域图上估计所有点间的测地线距离,测地线距离用点间的最短路径近似;第三,利用 MDS 计算低维嵌入.

LE 算法是基于谱图理论的方法,它将从数据集得到的图形拉普拉斯算子近似为流形上的拉普拉斯-贝尔特拉米算子.算法包括三个步骤:第一,确定每个对象的 k 个近邻点,构建邻域图;第二,为每条边选择一个权值,形成权值矩阵,权值可以用热核方程或简单的方法确定;第三,进行特征映射,利用拉普拉斯算子将权值矩阵转化为推广的特征值问题,计算特征值来得到低维表示.

LTSA 算法与前面所述流形学习算法不同的地方在于高维数据样本点的邻域选取标准不同,LTSA 算法中样本点的邻域是用其所在领域的切空间表示的,并且建立每一个点的邻域切空间,最后通过所有点的邻域切空间的排列建立起低维流形的全局坐标.LTSA 算法基于这样的理论:理想的低维嵌入同局部的投影坐标之间只相差一个仿射变换,并由此构造一个最小化重构误差,求解最小化重构误差问题可以转化成求解一个稀疏矩阵的特征值问题[12].LTSA 算法也是首先构建邻域图,然后通过一个优化函数计算 d 维仿射子空间,最后求得低维嵌入.

2 基于流形学习的微阵列数据分类方法

基因微阵列数据中每个样本含有几千甚至上万个基因,具有很高的维数,直接使用分类算法对这些高维数据进行分类一方面会造成分类精度不高,另一方面会降低分类算法的执行效率.基因微阵列数据本身可以看作是嵌入在高维空间中的低维流形,如果使用流形学习算法对基因微阵列数据进行降维,将其投影到低维空间中,提取出与分类类别相关的样本特征,无疑会提高算法的执行效率,而且会提高分类识别的效果.基于此提出了基于流形学习的微阵列数据分类模型,如图 1 所示.

在该模型中,首先利用流形学习算法对微阵列数据进行降维,然后对降维后数据利用分类算法进行分类.该模型是一个流形学习算法与分类算法相结合的一般模型,流形学习算法可使用 LLE、ISOMAP、LE 和 LTSA 等算法中任何一个,分类算法可使用 KNN、Naive Bayes、SVM 等算法中任何一种.通过流形学习将基因微阵列数据映射到低维空间中,再对降维后的数据进行分类,最终能达到相对较好的识别效果,并提高分类算法的执行效率.

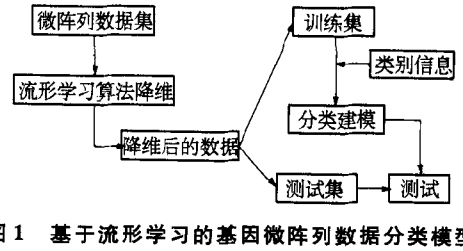

图 1 基于流形学习的基因微阵列数据分类模型

Fig. 1 Classified model of gene microarray data based on manifold learning

3 实验结果

为了验证所提出的分类模型的有效性,在白

血病数据集上进行实验. 该数据集由 38 个白血病的基因表达谱数据样本组成，其中每个样本包含 5000 个基因. 整个数据集包括急性髓细胞性白血病(AML)和急性淋巴细胞白血病(ALL)两种样本，其中 ALL 又可以分为 T 细胞(T_cell)和 B 细胞(B_cell)两个子类，因此整个数据集实际上分为三种样本，由 11 个急性髓细胞性白血病(AML)，19 个 B_cell 急性淋巴细胞白血病(ALL_B)和 8 个 T_cell (ALL_T)急性淋巴细胞白血病组成.

在实验中，将数据集分为训练集和测试集，其中训练集包括 10 个 ALL_B 细胞，4 个 ALL_T 细胞和 5 个 AML 细胞，剩余的样本作为测试集. 为了比较降维前后的分类效果，先在原始的白血病数据集上执行分类算法，然后再对白血病数据集利用流形学习算法进行降维后执行分类算法. 其中分类算法包括 K-NN、NB 和 SVM 算法，流形学习算法包括 ISOMAP、LLE、LE 和 LTSA 算法，在实验中对三种分类算法和四种流形学习算法进行组合得到 12 种分类器. 在一般的分类研究中，通常用识别率、查准率和召回率来评价分类算法的性能，在本文中也用这三个指标来评价基于流形学习的微阵列数据分类模型的性能.

在将微阵列数据映射到低维空间时，需要确定低维空间的维数，本文采用 ISOMAP 算法对白血病数据集进行降维，用残差曲线的拐点来近似数据集的本征维数，残差曲线如图 2 所示，其中邻域参数 k 为 3. 从图中可以看出残差曲线在维数为 3 时出现较明显的拐点，在本文中为了提高分类算法的精度，低维维数选择为 10. 对于 LLE 算法和 LE 算法邻域参数也选择为 3，对于 LTSA 算法邻域参数选择为 19，当邻域参数小于 19 时会出现奇异矩阵现象. 对于四个流形学习算法低维维数都选择为 10. 对于 K-NN 分类算法，根据先验知识 K 可以设置为 3. 实验结果如表 1 至表 3 所示.

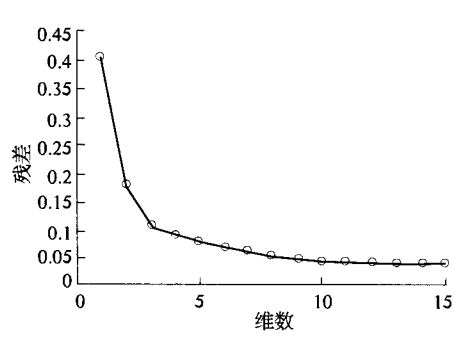

图 2　用 ISOMAP 算法得到的残差曲线

Fig. 2　The curve of the residual variance with ISOMAP

表 1　流形学习算法与 K-NN 算法结合的分类结果

Tab. 1　Results of combining manifold learning with K-NN

方法	识别率		查准率			召回率	
K-NN	0.7368	0.6429	1.0000	1.0000	1.0000	0.7500	0.3333
ISOMAP + K-NN	0.7895	0.6923	1.0000	1.0000	1.0000	1.0000	0.3333
LLE + K-NN	0.7895	0.6923	1.0000	1.0000	1.0000	0.7500	0.5000
LE + K-NN	0.8421	0.7500	1.0000	1.0000	1.0000	1.0000	0.5000
LTSA + K-NN	0.7895	0.6923	1.0000	1.0000	1.0000	1.0000	0.3333

表 2　流形学习算法与 NB 算法结合的分类结果

Tab. 2　Results of combining manifold learning with NB

方法	识别率		查准率			召回率	
NB	0.7895	0.6923	1.0000	1.0000	1.0000	0.5000	0.6667
ISOMAP + NB	0.9474	0.9000	1.0000	1.0000	1.0000	1.0000	0.8333
LLE + NB	0.9474	1.0000	1.0000	0.8571	0.8889	1.0000	1.0000
LE + NB	0.9474	1.0000	1.0000	0.8571	0.8889	1.0000	1.0000
LTSA + NB	0.8421	0.7500	1.0000	1.0000	1.0000	0.7500	0.6667

从实验结果可以看出，流形学习算法与分类方法结合后，与直接用原始数据进行分类相比，识别率、查准率、召回率都有了较大幅度的提高. 对于白血病数据集，直接用高维数据集进行分类，K-NN 分类算法的分类精度最低，SVM 的分类精度最高. 流形学习方法与分类算法结合后，总体上是流形学习算法与朴素贝叶斯法结合的效果最好，在四种流形学习方法中用 LE 算法得到的结果最好. 除了分类精度大幅提高外，利用降维后的数据进行分类也会极大地提高分类算法的执行效率.

表 3　流形学习算法与 SVM 结合的分类结果

Tab. 3　Results of combining manifold learning with SVM

方法	识别率		查准率			召回率	
SVM	0.8421	0.7500	1.0000	1.0000	1.0000	1.0000	0.5000
ISOMAP + SVM	0.8947	0.8182	1.0000	1.0000	1.0000	0.7500	0.8333
LLE + SVM	0.8947	1.0000	0.6667	1.0000	1.0000	1.0000	0.6667
LE + SVM	0.8947	0.9000	1.0000	1.0000	1.0000	1.0000	0.8333
LTSA + SVM	0.8421	0.7500	1.0000	1.0000	1.0000	0.7500	0.6667

4　结论

由于基因微阵列数据具有极的高维数，直接进行分类会降低分类算法的性能，因此降低数据的维

数是非常必要的.流形学习作为一种非线性降维方法,可以将高维数据有效地映射到低维空间中,发现其内在的流形结构.本文提出了一种基于流形学习的微阵列数据分类模型,首先利用流形学习算法将基因微阵列数据映射到低维空间中,然后再用降维后的数据进行分类.在实验中,讨论了四种流形学习算法 LLE、ISOMAP、LE 和 LTSA 算法与三种分类方法 K-NN、NB 和 SVM 结合的效果.通过实验得到以下结论:

(1)流形学习算法与分类方法结合后分类精度明显提高,同时有效提高了分类算法的执行效率.

(2)从实验结果可以得出,对于白血病数据集,ISOMAP、LLE 和 LE 算法与朴素贝叶斯法结合会取得最好的分类结果,流形学习方法与 SVM 的结合次之,流形学习方法与 K-NN 的结合最差.

参考文献:

[1] SLONIM D K, TAMAYO P, MESIROV J P, et al. Class prediction and discovery using gene expression data [C]. New York: ACM, 2000: 263 - 272.

[2] RAMASWAMY S, GOLUB T R. DNA microArrays in clinical oncology[J]. Journal of Clinical Oncology, 2002, 20(7): 1932 - 1941.

[3] KURAMOCHI M, KARYPIS G. Gene classification using expression profiles: A feasibility study [C]. New York: IEEE, 2000: 191 - 200.

[4] 李杰,唐降龙,王亚东,等.基因表达谱聚类 P 分类技术研究及展望[J].生物工程学报,2005,21(4):667 - 673.

[5] 于化龙,顾国昌,赵靖,等.基于 DNA 微阵列数据的癌症分类问题研究进展[J].计算机科学,2010,37(10):16 - 22,32.

[6] 王明怡,吴平,夏顺仁.基于人工神经网络集成的微阵列数据分类[J].浙江大学学报:工学版,2005,39(7):971 - 975.

[7] 陈磊,刘毅慧.基于 CART 算法的肺癌微阵列数据的分类[J].生物信息学,2011,9(3):229 - 234.

[8] STATNIKOV A, ALIFERIS C F, TSAMARDINOS I. A comprehensive evaluation of multicategory classification methods for gene expression cancer diagnosis[J]. Bioinformatics, 2005, 21(5): 631 - 643

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

[10] TENENBAUM J B, DESILVA V, LANGFORD J C. A global geometric framework for nonlinear dimensionality reduction[J]. Science, 2000, 290: 231 - 2323.

[11] BELKIN M, NIYOGI P. Laplacian eigenmaps and spectral techniques for embedding and clustering [C]. Cambridge: MIT Press, 2002: 585 - 591.

[12] ZHANG Zhen-yue, ZHA Hong-yuan. Principal manifolds and nonlinear dimensionality reduction by local tangent space alignment[J]. SIAM Journal of Scientific Computing, 2004, 26(1): 313 - 338.

A Classification Method Based on Manifold Learning for Gene Microarray Data

LI Qiang, SHI Lu-kui, LIU En-hai, WANG Ge

(School of Computer Science and Engineering, Hebei University of Technology, Tianjin 300401, China)

Abstract: Each sample in gene microarray data contains thousands or even tens of thousands of genes. It is necessary to reduce the dimension of the data before classifying them for obtaining better classified results. Manifold learning, as a nonlinear dimension reduction method, can discover the intrinsic laws hidden in the high dimensional data and has been widely applied in areas such as pattern recognition. A model combining manifold learning with classified algorithms was proposed to classify microarray data. In the model, the dimension of microarray data was firstly reduced with some manifold learning method. Then the data reduced the dimension were classified. In experiments, several manifold learning algorithms including LLE, ISOMAP, LE and LTSA are combined with three classified methods. And the results are compared with those from directly classifying high dimensional data. Experiments showed that the classification accuracy was great improved with the proposed model. Moreover, the execute efficiency of classification algorithms was also greatly increased.

Key words: manifold learning; classification; gene; microarray data