基于邻域粗糙集的加权 KNN 肿瘤基因表达谱分类算法①

陈智勤　(福建师范大学 数学与计算机科学学院 福建 福州 350007)

摘　要： 肿瘤亚型的准确判别对肿瘤的治疗具有重要意义,对肿瘤的不同亚型进行准确判别是当前生物信息学研究的重要课题.本文首先利用 Relief 算法排序基因并选出初始的肿瘤信息基因子集,然后利用向基于邻域粗糙集模型的向前属性约减算法 FARNeM 来计算加权基因集合,最后用加权 KNN 算法对肿瘤对这些数据进行分析, 从而发现有差异的基因表达。实验结果表明了上述方法的可行性和有效性。

关键词： 基因表达谱；肿瘤分类；邻域粗糙集；加权 K-NN 算法

Weighted KNN Algorithm for Tumor Gene Expression Profiles Classification Based on Neighborhood Rough Sets

CHEN Zhi-Qin (School of Mathematics and Computer Science, Fijian Normal University, Fuzhou 350007, China)

Abstract: The accurate identification of tumour subtypes in the treatment of tumors is important; the classification of different tumor subtypes has recently received a great deal of attention in the field of bioinformatics. The paper sorts genes using Relief algorithm and selects the initial subset of the genes of tumor information firstly. Then, calculates the weighted gene sets using the forward attribute reduction algorithm based on neighborhood rough set model. Then the weighted K-NN algorithm is used to analyze the data in order to detect differentially expressed genes. The results showed the feasibility and effectiveness of the method proposed in this paper.

Keywords: gene expression profiles; tumour classification; neighborhood rough sets; weighted K-NN algorithm

1　引言

DNA 微阵列技术的到来将对生物学和医学产生一场革命,微阵列实验正在生物学和医学研究中帮助研究人员解决越来越多的问题。随着大规模基因表达谱技术的推广,人们利用 DNA 芯片可以在一次实验中同时获得组织样本中成千上万个基因的表达数据。如何从基因表达谱数据中选取包含样本分类信息的特征基因，建立分类器，实现肿瘤的分型诊断是当前生物信息学研究的重要领域[1-4]。当前,对基因表达数据进行分类分析的主要方法还有人工神经网络、 遗传算法、支持向量机、贝叶斯和 K-近邻法等。

粗糙集理论[5,6]作为一种研究现实中各种获得信息的数学理论 ,主要是以集合的整体直接逼近的方式完成对不完整不确定信息前提下的知识推理过程。近年来，随着数据挖掘领域的兴起，粗糙集理论发展很快，应用更加广泛，已逐步扩大到基因表达谱数据挖掘和肿瘤分类等研究领域。文[7]提出了基于粗糙集理论的肿瘤样本分类方法；文[8]采用粗糙集方法来预测白血病，并在白血病数据集样本中发现 8 个与白血病有关的基因和 8 个分类信息规则；针对数据海量现状也有基于粗糙集理论的知识约简研究[9]与特征选择研究[10,11]。本文将邻域粗糙集模型与加权 KNN 算法相

① 基金项目:福建省自然科学基金(07J0016);福建省教育厅 B 类项目(JB09057)
收稿时间:2010-04-12;收到修改稿时间:2010-05-30

结合，提出了一种基于邻域粗糙集的加权 KNN 肿瘤基因表达谱分类算法。该算法可以有效的避免同样重要的依赖于所有属性的相似性度量引起的误导，可以克服“维数陷阱”问题。

2 Relief基因选择算法

Relief 算法作为一种属性重要性排序的机器学习算法在特征选取方面得到广泛的应用，其核心思想是以属性区分“相近”样本的能力作为评估属性重要性的标准，并由此给出属性的分类权重。Relief 算法首先对给定的一个样本找到与它距离最近的两个邻居：一个邻居来自与它相同类别的群体，另一个来自相异的类别的群体。然后在训练集中搜索某一样本近邻的过程是以两个样本之间的距离为标准进行的。

基于 Relief 算法来选择与肿瘤相关的信息基因的算法伪代码描述如下：

Relief 算法(Strn,F) //F 为待分析的属性集合，Strn 为训练样本集

Step1: Set weights vector W to zeros

//向量 W 中第 i 个元素对应于 F 中的第 i 个属性的分类权重

Step2: For i=1 to card(Sm)

//card(Sm)为样本集 Strn 中的样本数

Choose the i-th instance s in Strn

Find its nearest K Hits and nearest K Misses

// K>=1,当 K>1 时称为 Relief-A 算法

For j=1 to card (F)

W[g]:=W[g]-diff (g, Ri, H)/m+diff (g, Ri, M)/m

End

End

Step3:Return W //返回权重向量

其中 diff(g,s1,s2)用于计算基因 g 在样本 s1 和 s2 中的差异，定义为：diff(g, s1,s2)=|value(g,s1)-value(g,s2)|/(max(g)-min(g))

3 邻域粗糙集基础

邻域粗糙集模型是由胡清华在经典粗糙集理论模型的基础上提出，能够直接处理连续数据而不需要事先对其进行离散化处理的方法。由于在基因约简前不存在信息损失问题，因此选出的基因子集具有更强的分类能力。

定义 1. 给定样本集合 $U=\{x_1,x_2,\mathbf{L},x_n\}$，A 为属性集，C 是描述 U 的实数型特征集合，D 是决策属性集合，如果 C 生成论域 U 上的一簇邻域关系，则称 $NDS=\langle U,A=C\mathbf{U}D\rangle$ 为一个邻域决策系统。D 将 U 划分为 N 个等价类：$X_1,X_2,\mathbf{L},X_N,\forall B\subseteq C$, 定义决策 D 关于 B 的下近似和上近似分别为：

$$Lower(D,B)=\bigcup_{i=1}^{N}Lower(X_i,B)$$

$$Upper(D,B)=\bigcup_{i=1}^{N}Upper(X_i,B) \quad (1)$$

其中，$Lower(X,B)=\{x_i\,|\,d_B(x_i)\subseteq X,x_i\in U\}$；$Upper(X,B)=\{x_i\,|\,d_B(x_i)\,\mathbf{I}\,X\neq f,x_i\in U\}$，$d_B(x_i)$ 是由属性 B 和度量 Δ 生成的邻域信息粒子。

定义 2. 给定一个邻域决策系统 $NDS=\langle U,A=C\mathbf{U}D\rangle$，设 $\forall B\subseteq C$，那么决策属性 D 关于条件属性 B 的依赖度定义为：

$$g(D,B)=Card(Lower(D,B))/Card(U) \quad (2)$$

显然，$0\le g(D,B)\le 1$。

定义 3：给定一个邻域决策系统 $NDS=\langle U,A=C\mathbf{U}D\rangle,\forall B\subseteq C,\forall a\subseteq B$

如果 $g(D,B-a)<g(D,B)$，则称 a 关于 B 是必要的；否则，如果 $g(D,B-a)=g(D,B)$，则 a 是冗余的。如果都是必要的，则称 B 是独立的。如果 B 满足：$\forall a\in B,g(D,B-a)<g(D,B)$ 和 $g(D,B)=g(D,C)$

则称 B 为 C 的一个约简。若 $B_1,B_2,\mathbf{L},B_K$ 是此系统的全部约简，则称 $Core=\bigcap_{i=1}^{K}B_i$ 为系统的核。

对于肿瘤亚型分类，可形式化表示为 $NDS=\langle S,A=C\mathbf{U}D,V,f\rangle$ 这样的一个邻域决策表，其中 $S=\{s_1,s_2,\mathbf{L},s_m\}$ 是一个非空肿瘤样本集，称之为一个样本空间。$G=\{g_1,g_2,\mathbf{L},g_m\}$ 是一个非空基因子集，称之为条件属性。$D=\{L\}$ 是一个输出特征变量，称之为决策属性，L 表示样本所属类别的标记。V_a 表示属性 $a\in G\mathbf{U}D$ 的值域，f 是一个信息函数，可以表示为 $f:S\times(G\mathbf{U}D)\to V$，其中 $V=\bigcup_{a\in G\mathbf{U}D}V_a$。

基于邻域粗糙集模型的向前属性约简(forward attribute reduction based on neighborhood model,FARNeM)算法的伪代码如下：

输 入 ： $NTD=\langle S.A=G\cup D,V,f\rangle$ and neighborhood d // d is the threshold to control the size of the neighborhood

输出：red; //基因 Z 子集，即 G 的约简

1： $\forall a\in G$ ： computing neighborhood relation N；

2： $red=f$；

3： for each $a_i\in G-red$ Computing $SIG(a_i,D,red)=g(D,red\cup a_i)-g(D,red)$；

4： Selecting gene a_k satisfying $SIG(a_k,D,red)=\max_i(SIG(a_i,D,red))$ ；

5： if $SIG(a_k,D,red)>0$ $red=red\cup a_k$；

go to 3

else

return red;

算法结束

4 基因表达谱分类算法

K-近邻分类(K-nearest neighbor,K-NN)算法是一种建立在通过类比学习的算法，它根据测试样本在特征空间中 k 个最近邻样本中的多数样本的类别来进行分类，因此具有直观、无需先验统计知识等特点。然而，传统的 K-NN 算法选择的相似性度量通常是欧几里得距离的倒数，也就是说，两者距离越小，表示两者相似性越大，反之则相似性越小。由欧式距离的定义可见，这种距离通常涉及所有属性，且认为这些属性对距离的影响程度是等同的。同等重要的依赖于所有属性的相似性度量会引起误导。克服此问题的一般性措施就是对每一个属性增加一个特征权重参数，以便不同的属性在分类中起不同的作用。本文提出了一种基于邻域粗糙集的加权 KNN 肿瘤基因表达谱分类算法(如图 1 所示)，利用邻域粗糙集的向前属性约简来选择所有属性中的重要属性，给这些属性赋予更大的权重。对于基因表达谱矩阵中的两个样本矢量 $X=\{x_1,x_2,\mathbf{L},x_p\}$ 和 $X'=\{x_1',x_2',\mathbf{L},x_p'\}$，其中，p 是基因个数，$x_i$ 是第 i 个基因的表达值，那么它们之间的相似度量采用欧氏距离：

$$d(X,X')=\sqrt{\sum_{i=1,a_i\in Z}^{p}b(x_i-x_i')^2+\sum_{i=1,a_i\notin Z}^{p}(1-b)(x_i-x_i')^2} \quad (3)$$

其中，Z 为基于邻域粗糙集模型的向前属性约简算法得到的约简后的基因集合，$a_i\in Z$ 表示第 i 个基因属性属于约简后的基因集合，b 为重要属性的权重 $b>0.5$。

基于邻域粗糙集的加权 KNN 肿瘤基因表达谱分类算法的步骤如下：

Step1:采用 Relief 算法排序所有基因，然后选择前 d 个基因构成初始信息基因子集 G_d；

Step2:采用基因邻域粗糙集模型的前向属性约简算法 FARNeM 对信息基因集合 G_d 中的基因进行约简，进一步得到重要信息基因子集 G_o,其基数接近最小；

Step3:近似重要信息基因子集 G_o 作为加权 K-NN 分类方法的输入，对肿瘤样本集进行分类训练，训练后得到肿瘤分类模型。

Step4:采用测试集评估分类模型。

5 实验结果分析

本实验采用 Golub 等收集的急性白血病的基因表达谱数据集作为实验样本集，该数据集共含 72 个急性白血病样本，每个样本均含 7129 个基因的表达数据。该数据集共含 72 个急性白血病样本,每个样本均含 7129 个基因的表达数据. 其中 47 个样本被诊断为急性淋巴性白血病(ALL) ,25 个被诊断为急性骨髓性白血病(AML) 。如图 1 所示:

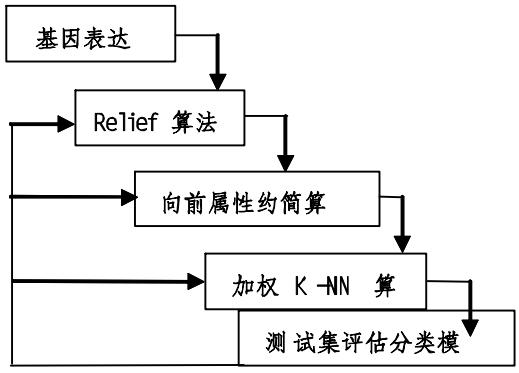

图 1 分类算法结构图

本实验主要分为 5 个阶段:

基因表达谱数据获取、用 Relief 算法进行数据预处理、利用向前属性约简算法选择重要信息基因子集、加权 K-NN 分类模型建立以及分类预测与评估。实验步骤，首先对该数据集采用 Relief 信息基因选择算法选择出与样本类别相关的基因，然后，采用加权 K-NN 算法进行肿瘤分类，最后对分类预测的准确性进行评估。为了提高分类精度的准确性与稳定性,本文采用 10-折交叉验证 (CV, Cross Validation)在训练集上进行样本识别,即将样本集分为 10 份 ,其中 9 份作为训练数据集 ,而另外的 1 份作为测试数据集。用测试集来验证所得分类器,循环 10 次,直到所有 10 份数据全部被测试 1 遍为止。每次循环得到一个识别率,取所有识别率的算术平均值作为是最终的识别率。

表 1 急性白血病实验结果比较

数据集与分类方法	基因数/描述	分类精度/%	参考文献
Relief+DCT+PCA+K-NN	PC=5	94.52	文[12]
FSC+SVM	-	94.1	文[13]
RFSC+SVM(RBF kernel)	16	100	文[14]
Relief+FARNeM+WKNN	$d=0.78, b=0.7$	98.02	本文
Relief+FARNeM+WKNN	$d=0.73, b=0.9$	100	本文

表 1 给出了本文的实验结果跟其他方法的比较，从表可以看出文献[12]的基因选择方法基与本文的基因选择方向相同,但是由于本文采用了加权 K-NN 而获得了更高的分类精度。本文方法在参数的情况下能获得与文献[14]一样 100%分类精度。从表 1 可以看出由于参数不同分类精度不一样，令参数从 0.15 至 0.95 均匀选取 10 个数字,参数从 0.55 至 0.95 均匀选取 5 个数字，两个参数组合进行 50 次分类计算得到的平均分类精度为 98.17%,比文献[12]和文献[13]的精度要高。以上实验表明将邻域粗糙集模型与加权 K-NN 算法结合应用于基于基因表达谱的肿瘤分类是一种可行的有效的方法。

6 结论

本文重点研究了基因表达数据分类问题 ,提出了一种基于邻域粗糙集的加权 KNN 肿瘤基因表达谱分类算法，该算法将邻域粗糙集模型与加权 KNN 算法相结合的一种监督学习技术。由于采用了邻域粗糙集模型的向前属性约简(FARNeM)算法来选取重要特征基因子集，在基因表达谱分类算法中给重要特征基因子集赋予更大的权重，这样可以有效的避免同样重要的依赖于所有属性的相似性度量引起的误导，可以克服“维数陷阱”问题。对急性白血病数据集进行了实验，实验结果表明该分类方法是确实可行的有效的。

参考文献

1 Golub TR, Slonim DK, Tamayo P, et al. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. Science, 1999,286(5439):531－537.

2 Brown MPS, Grundy WN, Lin D, et al. Knowledge-based analysis of microarray gene expression data by using support vector machines. Proc Nat'l Acad Sci, 2000,97(1):262－267.

3 Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science, 1999,286:531－537.

4 李颖新,阮晓钢. 基于支持向量机的肿瘤分类特征基因选取. 计算机研究与发展, 2005,42(10): 1796－1801.

5 PAWLAK Z. Rough sets. International Journal of Computer and Information Sciences, 1982,11(5):341－356.

6 PAWLAK Z. Rough sets: theoretical aspects of reasoning about data. Boston: Kluwer Academic Publishers, 1991.

7 Midelfart H, Komorowski J, Noresett K, et al.Learning rough set classifiers from gene expressions and clinical data. Fundamenta, 2002,53:155－183.

(下转第 16 页)

(上接第 89 页)

8 Fang JW, Grzymala-Busse JW. Leukemia prediction from gene expression data-a rough set approach. Annual Kansas City Area Life Sciences Research Day. Kansas City, MO. 2006.

9 王加阳,陈松乔,罗安. 粗集动态约简研究. 小型微型计算机系统, 2006,27(11):2056－2060.

10 Quafafou M, Boussoufm. Generalized rough sets based feature selection. Intelligent Data Analysis, 2000,4(1):3－17.

11 Han JC, Hu XH, Lin TY. Feature subset selection based on relative dependency of attributes: Proc of the 4th International Conference on Rough Sets and Current Trends in Computing. Berlin: Springer, 2004:176－185.

12 黄德双.基因表达谱数据挖掘方法研究.北京:科学出版社, 2009.

13 Furey TS, Cristianini N, Duffy N, et al. Support vector machine classification and validation of cancer tissue sample using microarray expression data. Bioinformatics, 2000(16):906－914.

14 史朝晖. SVM 算法及其在 HRRP 分类中的应用[硕士学位论文].西安:空军工程大学, 2005.