计算机科学2002Vol.29№.6

参差比较和遗传算法的中药识别分类器

Chinese Traditional Medicine Identification Classifier Using
Corresponding Relative Distance Computing and Genetic Algorithm

张丽新[1] 赵雁南[1] 蔡少青[2] 杨泽红[1] 刘宏宇[2] 王家钦[1]
(清华大学计算机科学与技术系 智能技术与系统国家重点实验室 北京100084)[1]
(北京大学医学部药学院 北京100083)[2]

Abstract The identification of Chinese traditional medicine is a difficult subject in pharmacology. The development of chemical measurement and pattern recognition make the chemical pattern recognition possible. In this paper a new chemical pattern recognition method is proposed, in which a simple method named corresponding relative distance computing is used to compute the distance between samples for Nearest-Neighbor (NN) classifier, and the genetic algorithm is used to optimize the parameters of NN classifier. With the proposed method in this paper, experiments are carried out on chromatogram data of Panax. The results indicate that the method can identify the medicine of different harvest time or habitat, furthermore, this method which combines the pattern matching, genetic algorithm and NN classifier is robust, accurate and easy to be implemented.

Keywords High performance liquid chromatogram, Pattern recognition, Corresponding relative distance computing, Genetic algorithm, Nearest-Neighbor classifier

1. 前言

中医药学是中国人民的宝贵财富,但常有中药替代品甚至伪劣品出现。传统的中药形态鉴定法和理化鉴定法耗时,且受主观影响大。近几十年来,色谱学和光谱学等化学测量手段的发展和计算机模式识别的成熟,使得中药的化学模式识别成为中药识别领域的研究热点[1]。

目前的中药化学模式识别方法主要有PCA(主成分分析法)、PRIMA法、神经网络、模糊聚类等[2~5],要从原信息中提取特征,构成特征向量,对这些特征向量进行处理。由于色谱图实验数据获取所需时间较长,因此一般样本量较少,不适于采用神经网络法,而且一般在采用神经网络方法前,需要采用PCA方法提取主成分,所需时间较长,信息失真可能性大。模糊聚类方法属于非监督分类方法,非监督方法的识别率一般比监督方法要低。

本文提出一种新的实用的化学模式识别方法,该方法采用一种简单直观的参差比较距离法来计算最近邻分类器中样本间的距离,同时采用遗传算法对数据预处理和样本间距离计算涉及到的参数进行优化。对不同收获时期和不同产地人参药品的色谱图数据进行中药识别难度较大,研究较少[1],应用本文提出的方法对此类数据进行分类实验,结果证明了这种方法的有效性。

2. 分类器总体设计

总体设计流程图如图1所示,分为训练和测试两个过程。训练过程完成分类器的设计。测试过程对设计好的分类器进行测试。训练过程如下:首先训练样本利用峰面积阈值参数T进行样本特征峰过滤,然后利用样本峰符合参数D计算样本间的距离,最后利用最近邻分类器进行分类,训练过程的分类验证方法为Leave-One-Out方法,分类结果作为遗传算法的适配值,由遗传算法对T和D参数进行优化。测试过程为:1)输入测试样本;2)进行样本特征峰过滤;3)计算与各训练样本的距离,将其分到和它距离最近的训练样本所属类别中去;4)输出测试样本的分类结果。

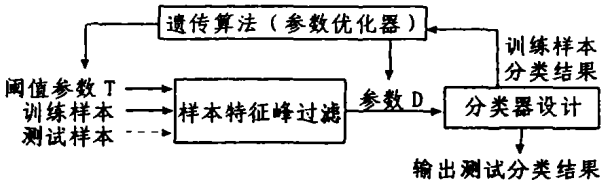

图1 分类器总体设计

3. 分类器各部分详细介绍

3.1 数据分析和数据结构定义

从惠普液相色谱系统HP化学工作站可以得到药品色谱图的峰个数,以及每个峰的保留时间、百分比峰面积等特征数据。处理后可得到如图2所示的色谱图简化模式,简化模式由若干峰组成,每个峰用峰保留时间、峰的百分比峰面积来描述。图2中横坐标为各峰出现的时间(保留时间),对应于一定的化学成分,纵坐标为峰的百分比峰面积(从峰开始出现到峰结束峰线的包络面积),表示该成分的百分含量。因此,每个峰的化学意义为该中药中含有某固定成分的百分比。

峰的百分比峰面积定义如下:

$$\text{某峰的百分比峰面积}=\frac{\text{该峰的峰面积}}{\text{该样品所有峰的峰面积和}}$$

采用百分比峰面积可以实现数据的归一化,消除了坐标尺度变化产生的影响。

描述每个样品的数据结构定义如下:

```
typedef struct {
int classtag;
int peeknum;
float peek[MaxPeekNum][2];
} PATTERN;
```

其中classtag表示该样本的类别,peeknum为该样本所含有

的最初峰数，peek[MaxPeekNum][2]描述样本峰信息，例如peek[1][1]表示第一个峰的保留时间，peek[1][2]表示第一个峰的百分比峰面积。

这样一个样品药材的模式定义展开为：类别，峰总数 n，(第一峰的保留时间，第一峰的百分比峰面积)，(第二峰的保留时间，第二峰的百分比峰面积)，…，(第 n 峰的保留时间，第 n 峰的百分比峰面积)。

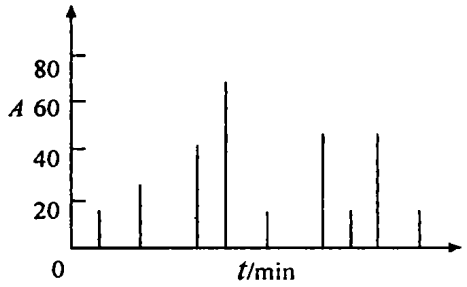

图2　药品色谱图的简化模式

保留时间的变异一般<0.1[6]，横坐标的范围一般为100左右，这样如果以0.1的间距进行离散化，特征的维数就近于1000维。这无疑是个高维问题，而且不同药品具有不同成分，所以无法统一降维。这给问题的处理带来很大的困难。为了解决这个困难，主成分分析和神经网络采取人为抽取特征的方法，即根据样本的特点，抽取某些固定位置(如大部分样本峰面积都较大的位置)的峰为样本属性，然后对这些峰数据进行主成分分析[2,4,5]。这种处理数据的方法主观性强，而且很难概括样本峰的真实分布。因此本文采用一个自适应确定的百分比峰面积阈值参数对样本的特征峰进行过滤，并相应地提出了参差比较距离计算方法来表征样本间的相似性。

3.2　参差比较距离计算方法

最近邻法是一种理论上较成熟直观易懂的分类方法，简单说，就是对于未知样本 x，比较其与所有已知样本的距离，并将 x 归到离它最近的样本所属类别中去[7]。

问题的关键在于距离的计算，与传统的欧式距离不同，根据中药色谱图数据的特点，本文提出一种动态参差比较距离计算法，所谓动态，是相对其他分类方法中常用的人为指定每个样本的代表峰数(如固定位置的10个峰)而言，具体处理分为两步：

首先采用阈值过滤法进行数据预处理，即设定样本峰过滤阈值参数 T，取那些百分比面积比它大的峰为样本的代表峰，这样每个样本所取峰数是不一样的，也许第一个样本用10个峰代表，而另一个样本要用20个峰，这就动态决定了每个样本所取峰数，滤掉了杂质成分，提高了后处理的效率。

然后计算两样本的距离，我们采用 D 参数作为不同样本中符合峰(两峰代表同一成分)的阈值。所谓符合峰，就是指不同样本的两峰的保留时间相差在 D 以内的，也即代表同一成分的峰。其面积差代表符合峰的距离。如果样本1中峰 I 在样本2中没有对应的符合峰，则样本2中该峰的面积为0，那么就可以计算它们对应的面积差，两样本所有符合峰面积差绝对值的和就代表两样本之间的距离。在样本间距离计算方法确定后，就可以直接采用最近邻法进行分类判别了。参差比较距离法分类流程图如图3所示，参差比较距离计算方法如图4所示。为了确定 T 和 D 参数的合理取值，我们采用了遗传算法进行优化，有关遗传算法优化参数部分将在下一节介绍。

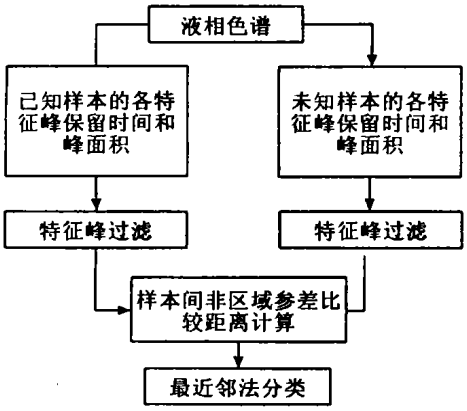

图3　分类流程图

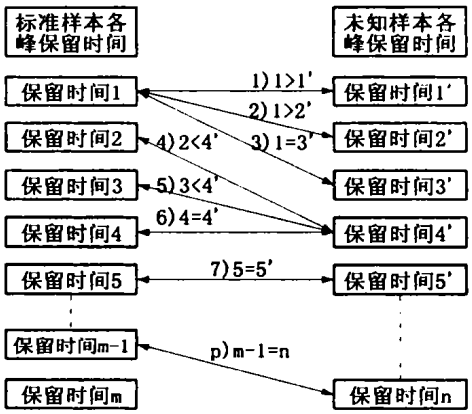

图4　参差比较样本距离计算示意图

3.3　遗传算法参数优化方法

分类器的设计过程涉及到两个参数，样本峰过滤阈值参数 T 和样本峰符合阈值参数 D。T 的作用是滤除中药里众多的杂质成分，突出主要成分的作用。如果选得过低，则起不到去除噪声的作用，如果取得过高，则会丢失有用的化学成分，不同 T 的效果如图5所示。保留时间变异系数也很重要，决定了当两峰保留时间相差超过多少时，认为两者不是同一成分。本文采用遗产算法优化上述两参数，采用训练过程的分类正确率做为评价函数，借鉴了 Wrapper 方法的思想[8]，取得了很好的效果。用遗传算法优化参数的过程如图6所示，有关遗传算法的细节，可以参见文[9]。

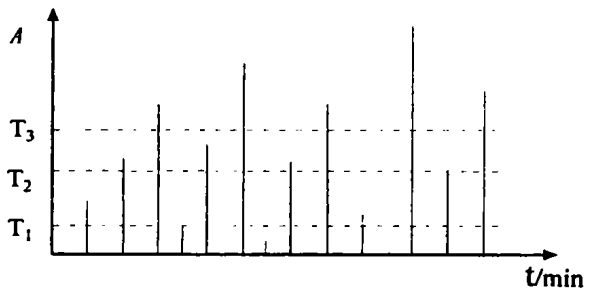

图5　不同阈值对样本峰的过滤影响(T 表示阈值)

4. 结果与讨论

实验数据为不同产地，不同收获日期的人参样品 HPLC 给出的数据，由北京大学医学院提供。不同产地和不同收获日

期的药品作为不同类别，可分为8类。有27个训练样本和23个测试样本，训练样本集和测试样本集为医学院先后提供的两批数据。

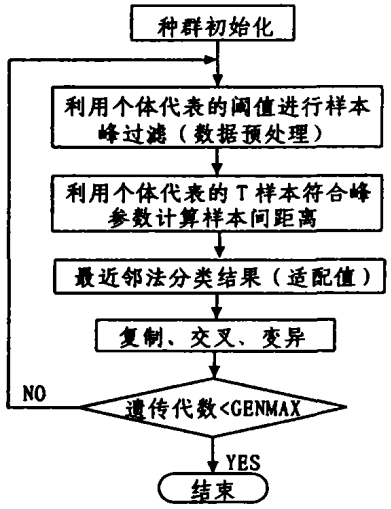

图6 遗传算法优化参数流程图

采用 Leave-One-Out 的验证方法[10]对训练分类器进行验证，即每次从训练样本集中取一个未验证过的样本做为测试样本，其他的样本做为训练样本。这样依次下去，就可以得到训练集中所有样本的最优的测试结果，从而可以得到一个训练样本分类正确率。

$$\text{训练样本分类正确率}=\frac{\text{所有训练样本中被正确分类的数目}}{\text{所有训练样本的数目}}$$

最后用测试集样本对分类器进行测试，可以得到测试样本分类正确率。

实验结果如表1所示。

表1 样本按不同产地不同日期的分类结果

类别	训练样本数	训练样本分类正确率	测试样本数	测试样本分类正确率
pa0828	3	100%	3	100%
pa0911	3	100%	3	100%
pa0925	3	100%	3	100%
xj0820	6	100%	2	100%
xj0905	3	100%	3	100%
xj1005	3	100%	3	100%
zjpa98	3	100%	3	66.7%
zjxj98	3	100%	3	100%
总分类正确率	27	100%	23	95.65%

结果分析：

1）虽然样本数和一般的分类相比较少，但是由于此领域研究取数据耗时间长，因此样本一般都不多[2~6]，而且由于本方法并未采用相关的统计信息，只是采用了最近邻法进行分类，而最近邻法的分类正确率通常随着样本数的增加而增加，因此可以预见当样本数增加时，分类正确率不会有很多降低。

2）关于分类错误样本的详细说明：zjpa98的一个测试样本被错误分类到 zjxj98，从色谱图上看，这两类样本的确很相似，当样本数足够多时，可以采用多级分类器的方法解决此问题，即先进行粗分类（如选取较大峰面积过滤阈值 T），这时可将不同类的相似样本分到一个粗类里，然后再进行细致的分类（如选取较小的峰面积过滤阈值 T）。

3）用遗传算法收敛很快，本文采用了20代，每代30个体的规模很快得到结果，在第一代，就可以得到最优结果，在第三代就明显收敛。而且遗传算法的评价函数为训练样本集的分类正确率，可以保证得到较高的训练样本分类准确率，经过多次验证，效果都很好。

4）所见文献提到的中药识别，都未细致到对不同收获日期不同产地的同属样品进行鉴别。本文提出的方法可以进行此种鉴别，并能得到较好结果，那么可以推测这种方法对不同种属药品进行鉴别可以得到更好的识别结果。

参 考 文 献

1 罗旭，毕开顺，等. 中药质量化学模式识别研究的进展. 药学学报，1993，28(12)：936～940

2 乔延江，王玺，等. 人工神经网络在中药蟾酥化学模式识别特征提取中的应用. 药学学报，1995，30(9)：698～701

3 吴玉柱，罗旭，等. 中药大黄质量的化学模式识别. 药学学报，1991，26(2)：132～138

4 王秀坤，李家实，等. 苦参质量的化学模式识别. 中国中药杂志，1996，21(4)

5 刘谦光，陈战国，等. 西洋参质量的化学模式识别. 中草药，1999，30(11)

6 杨维稼，杨健，等. 挥发油未知生药的气相色谱—计算机检索. 中日友好医院学报，1996，10(4)

7 边肇祺，张学工，等. 模式识别. 北京：清华大学出版社，2000，第二版

8 Kohavi R，John G H. The Wrapper Approach，In Feature Selection for Knowledge Discovery and Data Mining，H. Liu & H. Motoda(eds.)，Kluwer Academic Publishers，pp 33～50

9 陈国良. 遗传算法及其应用. 北京：人民邮电出版社，1996

10 Fukunaga K. Introduction to statistical pattern recognition. Boston：Academic Press，c1990，2nd ed.

（上接第97页）

$$=\max_{y\in V}\min\{R(u,y),\max_{z\in W}\min\{S(y,z),X(z)\}$$

$$=\max_{y\in V}\max_{z\in W}\min\{R(u,y),\min\{S(y,z),X(z)\}\}$$

$$=\max_{z\in W}\min\{\max_{y\in V}\min\{R(u,y),S(y,z)\},X(z)\}$$

$$=\max_{z\in W}\min\{R\circ S(u,z),X(z)\}=H_TX(u)$$

故 $H_T=H_R\circ H_S$。再证(1)：由定理5及(2)的结果可得：

$$\forall\ X\in F(W)(L_R\circ L_S)\ X=L_R(-H_S(-X))$$

$$=-H_R(H_S(-X))=-H_T(-X)=L_TX$$

故 $L_T=L_R\circ L_S$。

参 考 文 献

1 Pawlak Z. Rough sets[J]. International Journal of Computer and Information Sciences，1982，11(5)：205～218

2 张文修，吴伟志，梁吉业，等. 粗糙集的理论与方法[M]. 北京：科学出版社，2001

3 曾黄麟. 粗集理论及其应用[M]. 重庆：重庆大学出版社，1998

4 王珏，苗夺谦，周育健. 关于 Rough set 理论与应用的综述[J]. 模式识别与人工智能，1996，9(2)：337～344

5 祝峰，何华灿. 粗集的公理化[J]. 计算机学报，2000，23(3)：330～333

6 Yao Y Y. Constructive and algebraic methods of the theory of rough sets[J]. Information Sciences，1998，109 (1)：1～47

7 Thiele H. On axiomatic characterizations of crisp approximation operators[J]. Information Sciences，2000，129(1)：221～226

8 Morsi N N，Yakout M M. Axioms for fuzzy rough sets[J]. Fuzzy Sets and Systems，1998，100(2)：327～342