决策树和判别分析在分类问题中的比较研究

肖 娟，唐邵玲

（湖南师范大学 数学与计算机科学学院，湖南 长沙 410081）

摘 要：就分类的两种常见方法（统计中的判别分析和数据挖掘中的决策树方法）作对比研究，并用具体的事例数据分别针对预测准确率、速度、鲁棒性、易理解性等几方面进行研究，比较得出这两种方法的优缺点。

关键词：判别分析；决策树；预测；鲁棒

中图分类号：TP274+.3 **文献标识码**：A **文章编号**：1673-9833(2008)03-0027-05

Comparison Research on Decision Tree and Distinction Analysis in Classified Questions

Xiao Juan, Tang Shaoling

（School of Mathematics and Computer Science，Hunan Normal University，Changsha 410081，China）

Abstract: There are two common methods of distinction analysis in the statistics and decision tree method in the data mining for comparison research. It can obtain some good and bad points of these two methods separately according to aspects such as forecasting accuracy rate, the speed, robustness, understanding easily and so on with the concrete instance data.

Key words: distinction analysis; decision tree; forecast; robustness

0 引言

在生产、科研和日常生活中经常需要根据观测到的数据资料，对所研究的对象进行分类。例如，在经济学中，根据人均国民收入、人均工农业产值、人均消费水平等多种指标来判定一个国家的经济发展程度所属类型；在市场预测中，根据以往调查所得的种种指标，判别下季度产品是畅销、一般或是滞销；在地质勘探中，根据岩石标本的多种特性来判别地层的地质年代，由采样分析出的多种成份来判别此地是有矿或无矿，是铜矿还是铁矿等；在油田开发中，根据钻井的电测或化验数据，判别是否遇到油层、水层、干层或油水混合层等。总之，在实际情况中需要判别的问题几乎到处可见。

1 判别分析模型

判别分析是判别样品所属类别的一种统计方法，其用途之广可以和回归分析相媲美。判别分析通常都要设法建立一个判别函数，利用此函数来判断。

判别函数的一般形式如下：

$$Y = a_1x_1 + a_2x_2 - \cdots + a_cx_a,$$

其中，Y为判别指标，根据所用方法的不同，可能是概率，也可能是坐标值或分值；x_1，x_2等为反映研究对象特征的变量；a_1，a_2等为各变量的系数，也称为判别系数。为了建立该函数必须使用一个训练样本，所谓训练样本就是已知实际分类并且各指标的观测值也已经测得的样本，它对判别函数的建立很重要，如果中间出现一例错分，就会导致判别函数的判别效果大大降低。

判别分析常用的判别方法可以分为参数法和非参数法，也可以根据资料的性质分为定性资料的判别分析和定量资料的判别分析。常用的判别法有[1]：最大似然法、距离判别法、Fisher判别法和贝叶斯（Bayes）判别法。

收稿日期：2008-01-18

作者简介：肖 娟（1982-），女，江西新余人，硕士研究生，主要研究方向为金融统计；
唐邵玲（1963-），女，湖南邵阳人，湖南师范大学副教授，主要研究方向为金融统计.

2 决策树模型及算法

所谓决策树就是一个类似流程图的树型结构，其中树的每个内部结点代表对一个属性（取值）的测试，其分支就代表测试的每个结果；而树的每个叶结点就代表一个类别，树的最高层结点就是根结点。为了对未知数据对象进行分类识别，可以根据决策树的结构对数据集中的属性值进行测试，从决策树的根结点到叶结点的一条路径就形成了对相应对象的类别预测，决策树可以很容易转换为分类规则。构造决策树算法其实是一个贪心算法，当构造决策树时，有许多由数据集中噪声或异常数据所产生的分支。树枝修剪就是识别并消除这类分支，以帮助改善对未知对象分类的准确性。

下面给出学习构造决策树的一个基本归纳算法，它是著名决策树算法ID3的一个基本版本，采用自上而下、分而制之的递归方式来构造一个决策树。

决策树算法[2, 3]（Generate decision tree）：

//根据给定数据集产生一个决策树

输入：训练样本，各属性均取离散数值，可供归纳的候选属性集为：*attribute_list*

输出：决策树

处理流程：

1）创建一个结点 N；

2）若该结点中的所有样本均为同一类别 C，则

//开始根结点对应所有的训练样本

3）返回 N 作为一个叶结点并标志为类别 C；

4）若 *attribute_list* 为空，则

5）返回 N 作为一个叶结点并标记为该结点所含样本中类别个数最多的类别；

6）从 *attribute_list* 选择一个信息增益最大的属性 *attribute_list*；

7）并将结点 N 标记为 *attribute_list*；

8）对于 *attribute_list* 中的每一个已知取值 a_i，准备划分结点 N 所包含的样本集；

9）根据 *attribute_list*=a_i 条件，从结点 N 产生相应的一个分支，以表示该测试条件；

10）设 S_i 为 *attribute_list*=a_i 条件所获得的样本集合；

11）若 S_i 为空，则将相应叶结点标记为该结点所含样本中类别个数最多的类别；

12）否则将相应叶结点标志为Generate decision tree（S_i, *attribute_list*−*test_attribute*）返回值。

3 事例分析

下面分别用2种方法对同一组数据进行判别分类操作，数据来源于一家银行关于借款客户还款情况的历史数据库中的一部分，数据中的 *Credit rating* 是信用分类变量，用0表示 *Bad* 类，用1表示 *Good* 类；其他变量分别是年龄、收入、每个客户拥有的信用卡数量及受教育程度等。首先用判别分析[4]（Discrimnant 过程）对这组数据处理的结果如表1。

表1 逐步判别分析结果

Tab. 1 The result of stepwise discrimination

Step	Entered	Wilks' Lambda							
		Statistic	df1	df2	df3	Exact F			
						Statistic	df1	df2	Sig.
1	*Income level*	0.734	1	1	2 462	894.339	1	2 462	0.000
2	*Number of credit cards*	0.633	2	1	2 462	712.379	2	2 462	0.000
3	*Age*	0.569	3	1	2 462	622.130	3	2 462	0.000

逐步判别分析运行记录，可见第一步是将 *Income level* 这个变量纳入分析过程，第二步是将 *Number of credit cards* 这个变量纳入分析过程，第三步则是将 *Age* 纳入分析过程，右侧则给出了Wilks'Lambda检验的具体结果。三步的检验结果都是拒绝原假设，说明这三步分别纳入判别函数的变量对正确判断分类都是起作用的。

表2给出了一个判别函数中各个变量的标准化系数，可用来判断函数主要受哪些变量的影响较大。同时，知道了该系数就可以写出标准化的判别函数式。本例的判别函数式如下：

$$Credit\_rating = 0.487\,Age + 0.708\,Income\ level - 0.558\,Number\ of\ credit\ cards。 \quad (1)$$

表2 由SPSS15.0得到判别函数的标准化系数表

Tab. 2 Standardized coefficients of canonical discriminated function from SPSS15.0

Independent Variables（自变量）	Function 1
Age	0.487
Income level	0.708
Number of credit cards	−0.558

实际上式（1）计算的是各观测在各个维度上的坐标值，则可以通过这个函数式来计算出各观测的具体空间位置。结果图形如图1、图2。

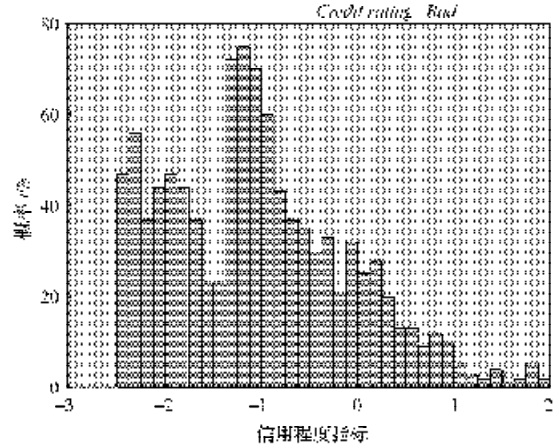

图 1　对应 Bad 类的标准判别函数图

Fig. 1　Stardardized canonical discriminant function of corresponding *Bad* class

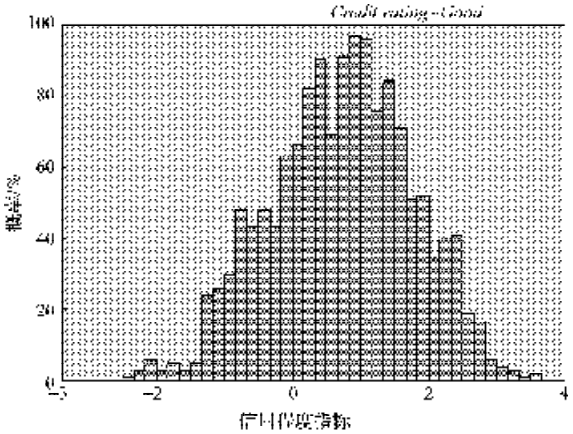

图 2　对应 Good 类的标准判别函数图

Fig. 2　Stardardized canonical discriminant function of corresponding *Good* class

接着给出相应 Fisher 判别函数的系数如表 3。

表 3　Fisher 判别函数的标准化系数表

Tab. 3　Standardized coefficients of Fisher canonical discriminated function

Independent Variables	*Credit rating*	
（自变量）	*Bad*	*Good*
Age	0.455	0.561
Income level	4.103	6.108
Number of credit cards	9.897	7.586
Constant	−20.299	−24.016

由此写出判别函数式为：

$$Bad:\ Y=-20.299+0.455Age+4.103Income\ level-9.897Number\ of\ credit\ cards,$$

$$Good:\ Y=-24.016+0.561Age-6.108Income\ level+7.586Number\ of\ credit\ cards.$$

可以用上面的两个判别式直接计算新观测记录属于各类的评分，得分最高的一类就是该观测记录相应的类别。例如对第一条记录可以计算出结果为：

$$Bad:\ Y=-20.299+0.455\times36.22+4.103\times2+9.897\times2=24.10866,$$

$$Good:\ Y=-24.016\ \ 0.561\times36.22\ 6.108\times2+7.586\times2=23.69142,$$

故把它归为 *Bad* 类。

再将用判别式预测出来的结果和真实结果作对比，得出预测结果的准确率如表 4。

表 4　判别分析预测结果的准确率情况表

Tab. 4　Accuracy condition table for discriminant prediction result

	Credit rating		*Predicted Group Membership*		*Total*
			Bad	*Good*	
Original	Count	*Bad*	828	192	1 020
		Good	298	1 146	1 444
	/%	*Bad*	81.2	18.8	100.0
		Good	20.6	79.4	100.0
Cross-validated	Count	*Bad*	828	192	1 020
		Good	298	1 146	1 444
	/%	*Bad*	81.2	18.8	100.0
		Good	20.6	79.4	100.0

由表 4 可以看出，将 *Bad* 类判断成 *Good* 类的有 192 条记录，正确率为 81.2 %；而将 *Good* 类判断成 *Bad* 类的有 298 条记录，正确率为 79.4 %；且整个判别分析的正确率为 80.1 %，整个程序运行时间为 1.74 秒。

最后用决策树的方法在 SPSS15.0 中做分析，需要说明的是输入到决策树的属性变量的个数和名称，包括 *Car loans, Education, Number of credit cards, Age, Income level*，Dependent Variable 则是样本记录中需要分类的分类变量。产生决策树的方法为 CHAID 算法，真正在产生决策树时用到的变量在 Results 中的 Independent Variable（即：*Number of credit cards, Age, Income level*）中给出，并且给出了节点的个数为 10，生成的分支终端个数为 6，即可以用 6 条规则把数据所给的记录进行分类，产生决策树的图形如图 3。从图 3 中很容易看出一些统计信息：如从根节点中可以看出，总共要分的类有 2 个，其中 *Bad* 类的分类记录有 1 020 条，占整个数据的 41.4 %；*Good* 类的分类记录有 1 444 条，占整个数据的 58.6 %。

从上述的图表中能很快地得到分类的规则：

1）If *Income level*<=*Low* Then *Credit rating=Bad*

2）If *Income level=Medium* And *Number of credit cards* <=5 Then *Credit rating=Good*

3） If *Income level=Medium* And *Number of credit cards*> =5 And *Age*<=28.08 Then *Credit rating=Bad*

4） If *Income level=Medium* And *Number of credit cards*>=5 And *Age*>28.08 Then *Credit rating=Good*

5）If *Income level=High* And *Number of credit cards*> =5 Then *Credit rating=Good*

6）If *Income level=High* And *Number of credit cards* <5 Then *Credit rating=Good*

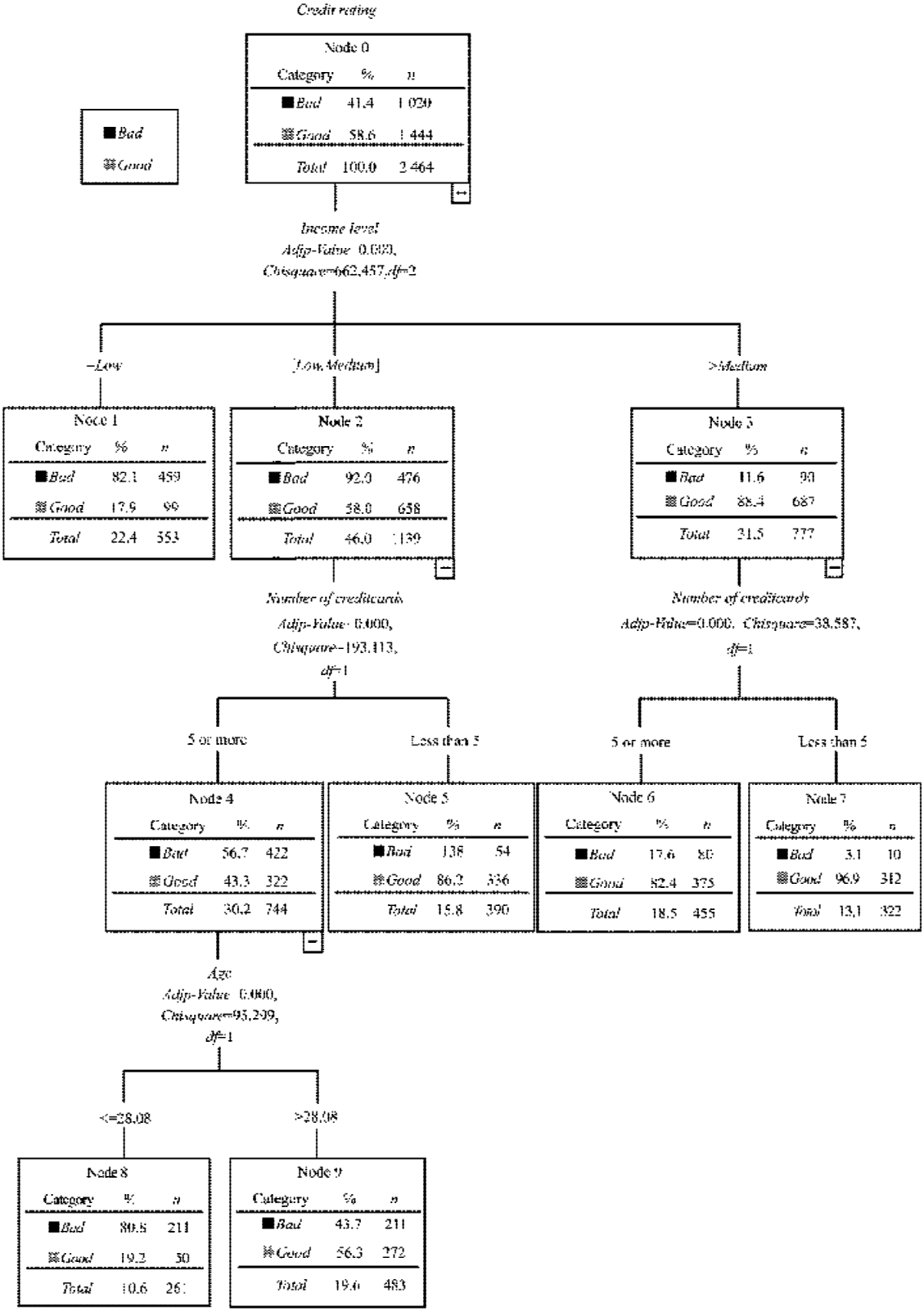

图 3 决策树算法的分析结果

Fig. 3 The result of decision tree

由上述 6 个分类的规则能对数据进行分类，例如对第一条记录：他的收入水平为 *Medium*，*Number of credit cards*>=5, *Age*>28.08，由上面的规则4不难发现这个人的信用水平应该是属于 *Good* 类。

用上述 6 个分类规则对样本进行分类后与真实的分类进行对比，发现把本来是 *Bad* 类的记录分成了 *Good*类的有355例，正确率为65.2 %；而把原来是*Good* 类错判为 *Bad*类的有 149 例，正确率为 89.7 %。整个决

策树算法的正确率为79.5 %，整个程序的运行时间为1.04秒。

表5　决策树算法预测结果的准确率情况表

Tab. 5　Accuracy condion table for prediction result of decision tree

Observed	Predicted		
	Bad	*Good*	*Percent Correct*
Bad	665	355	65.2%
Good	149	1 295	89.7%
Overall Percentage	33.0%	67.0%	79.5%

注：产生决策树的方法为CHAID算法，因变量为*Credit rating*。

4　结语

从决策树和判别分析两种方法对同组数据进行分析比较发现，这两种方法各有优势：

与其它分类方法相比，判别分析具有最小的错误率，但实际上由于其所依据的类别独立性假设和缺乏某些数据的准确概率分布，使得判别分析的准确率受到影响。从这个事例中可以看出在整体错判率上决策树的错判率为20.5%（100%-79.5%）较高，而判别分析的错判率为19.9%（100%-80.1%）较低，两者相差并不是很大。但是把*Good*类错判为*Bad*类的错判率上，决策树的错判率明显较低，这说明具体应用时，要看读者关心的是哪种错判率，可以选用不同的方法。

在计算速度上，很明显基于数据挖掘的决策树算法的速度（1.04秒）比判别分析的速度（1.74秒）要快很多，这就是为什么决策树归纳算法被广泛应用到许多进行分类识别的应用领域的原因。因为这类算法无需相关领域的知识，归纳学习和分类识别的操作处理速度都相当快，而对于具有细长条分布性质的数据集合来说，决策树归纳算法相应的分类准确率是相当高的，而判别分析的计算速度则没有决策树算法的速度快，因为它是完全基于概率统计的思想，需要的运算量都是比较大的，特别是关于大型数据的分类问题，判别分析对计算机的运算能力要求较高。

在数据的鲁棒性上，即在数据带有噪声和有数据缺失的情况下，决策树算法比判别分析正确预测能力强（见参考文献[5，6]）。

在方法的可扩展性和泛化上，决策树算法都要优于判别分析，决策树归纳方法可以与数据仓库技术结合到一起进行数据挖掘和分类工作，并且它还能利用多维数据立方存储分析基于不同抽象细度的数据，而判别分析则没有这样的性质。

没有一个分类方法在对所有数据集上进行分类学习均是最优的，选择一个分类方法需要从其预测准确性、训练时间、可理解性和可扩展性等诸多方面加以综合考虑。有关的研究表明：许多分类算法都是非常类似或接近的，它们间的差距在统计上几乎都是微不足道的；但它们的学习训练时间却有着较大的差别。一般而言，统计分类方法比多数决策树归纳学习方法需要更多的计算时间。

参考文献：

[1] 于秀林，任雪松. 多元统计分析[M]. 北京：中国统计出版社，2002.

[2] 邓纳姆(Dunham M H). 数据挖掘教程——世界著名计算机教材精选[M]. 郭崇慧，田凤占，靳晓明译. 北京：清华大学出版社，2005.

[3] （加）韩家炜，堪　博. 数据挖掘概念与技术[M]. 2版. 范明，孟小峰译. 北京：机械工业出版社，2007.

[4] 张文彤. 世界优秀统计工具SPSS11.0统计分析教程(高级篇)[M]. 北京：北京希望电子出版社，2002.

[5] （美）Trevor Hastie, Robert Tibshirani, Jerome Friedman. 数据挖掘、推理与预测[M]. 范　明，柴玉梅，咎红英译. 北京：电子工业出版社，2004.

[6] 梁　循. 数据挖掘算法与应用[M]. 北京：北京大学出版社，2006.

（责任编辑：罗立宇）