长沙通信职业技术学院学报
第9卷第4期 Journal of Changsha Telecommunications and Technology Vocational College Vol.9 No.4
2010年12月 Dec.2010

数据挖掘中两种简单分类算法的比较

王一夫，许 鹏，杨小林，韩 宇
（湖南师范大学数学与计算机科学学院，湖南长沙 41008）

【摘 要】分类是一类重要的数据挖掘问题，它的一般过程是先根据样本数据利用一定的分类算法得到分类规则，再依据该规则对新的数据进行类别的划分。文章详细介绍了两种简单但有效的分类方法：基于最小二乘法的线性分类器和k-最近邻分类器。通过对这两种分类器的比较，发现线性分类器计算简便、拟合具有低方差，适合处理类别之间相互重叠的区域比较小的数据。KNN分类器分类灵活，拟合偏差比较小，由于计算量比较大，该算法更适合于类别界限不是很明显，数据之间交叉或重叠比较多的数据集。

【关键词】最小二乘；最近邻；分类
【doi：10.3969/j.issn.1671-9581.2010.04.006】
【中图分类号】TP311.12 【文献标识码】A 【文章编号】1671-9581(2010)04-0022-04

Comparison of two simple classification algorithm in data mining

WANG Yi-fu, XU Peng, YANG Xiao-lin, HAN Yu
(Mathematics and Computer Science College, Hu´nan Normal University, Changsha,Hu´nan China 410081)

Abstract: Classification is an important question in data mining. Its general procedure is to obtain the classification rules according to the classification algorithm from the sample data firstly, then categorize the new data according to the classification rules. The author introduces two simple but effective classification algorithms in this paper: the linear classifier based on the least squares method and k-nearest neighbor classifier. Through comparison of these two classifiers, we draw the conclusion that the linear classifier has little computation, low variance and is suitable for handling with the data with small overlaps. KNN classifiers is more flexible, unbiased, but due to large computation, the algorithm is suitable for dealing with the data with relatively more overlaps.

Keywords: least square; nearest neighbor; classification

随着数据库应用的不断深化，数据库的规模急剧膨胀，数据挖掘已成为当今研究的热点。特别是其中的分类问题，由于其使用的广泛性，现已引起了越来越多的关注。分类技术在很多领域都有应用，例如可以通过客户分类构造一个分类模型来对银行贷款进行风险评估。采用数据挖掘中的分类技术，可以将客户分成不同的类别，比如呼叫中心设计时可以分为：呼叫频繁的客户、偶然大量呼叫的客户、稳定呼叫的客户等，帮助呼叫中心寻找出这些不同种类客户之间的特征，这样的分类模型可以让用户了解不同行为类别客户的分布特征。再比如文献检索和搜索引擎中的自动文本分类，安全领域中的入侵检测都是基于分类技术的实际应用。

在数据挖掘中[1]，输入数据，或称训练集，是一条条的数据库记录组成的。每一条记录包含若干条属性，组成一个特征向量。训练集的每条记录还有一个特定的类标签与之对应。分类的过程是根据样本数据利用一定的分类算法得到分类规则，新的数据过来就依据该规则进行类别的划分。分类的主要流程包含下面两个部分：

1） 训练：训练集→特征选取→训练→分类器
2） 分类：新样本→特征选取→分类→判决

[收稿日期] 2010-09-06
[作者简介] 王一夫（1970-），男，湖南衡阳人，副教授，博士，研究方向：计算神经科学、信号处理。
[基金项目] 湖南省科技厅项目（2009GK3014）和湖南省教育厅项目（09c636）资助。

在机器学习、统计学和神经网络等领域的研究人员已经提出了许多具体的分类预测方法。本文详细讨论了两种简单但有效的分类方法：基于最小二乘法的线性分类器和 k-最近邻分类器。通过对这两种分类器的比较，我们发现线性分类器对结构做了大量假定，并产生稳定但可能不精确地分类。k-最近邻对结构作了适度的假定，其预测常常是精确的，但可能不稳定。

1　线性分类器

在过去的 30 年中，线性模型一直是统计学的主要支柱，并且现在依然是我们最重要的工具之一。线性二分类器是使用线性模型来将数据分成两个类别，是一种最简单的分类器[2]。

给定一个输入向量 $X=(X_1,X_2,...,X_p)$，通过以下模型来预测输出 Y：

$$\hat{Y}=\hat{\beta}_0+\sum_{j=1}^{p}\hat{\beta}_j X_j \qquad (1)$$

其中 $\hat{\beta}_0$ 是截距，在机器学习中也称为偏置。通常，在 X 中包含一个常数变量 1，这样，向量形式的线性模型可以写成内积：

$$\hat{Y}=X^T\hat{\beta} \qquad (2)$$

其中，X^T 表示向量或矩阵的转置。这里对单个输出建模，因此 $\hat{Y}$ 是标量。一般来说，$\hat{Y}$ 可以是 K 向量。在 $(p+1)$ 维输入-输出空间中，$(X,\hat{Y})$ 表示一个超平面。如果 X 中包含常量，则超平面包含原点，是一个子空间。

为了用线性模型拟合训练数据集，我们必须估计模型中的参数，最常见的参数估计方法是最小二乘法。在这种方法下，我们选择系数 β，使得残差平方和最小[3]：

$$RSS(\beta)=\sum_{i=1}^{N}(y_i-x_i^T\beta)^2 \qquad (3)$$

$RSS(\beta)$ 是参数的二次函数，因此极小值总是存在，但可能不唯一。（3）用矩阵形式可以表示为：

$$RSS(\beta)=(y-X\beta)^T(y-X\beta) \qquad (4)$$

其中，X 是 $N\times p$ 的矩阵，每行是一个输入向量，而 y 是训练数据集中输出的 N 向量。关于微分，我们得到标准方程：

$$X^T(y-X\beta)=0 \qquad (5)$$

如果 X^TX 是非奇异的，则唯一解由下式给出：

$$\hat{\beta}=(X^TX)^{-1}X^Ty \qquad (6)$$

并且第 i 个输入 x_i 的拟合值为 $\hat{y}_i=\hat{y}(x_i)=x_i^T\hat{\beta}$。在任意输入 x_0 上，预测是 $\hat{y}(x_0)=x_0^T\hat{\beta}$。整个拟合面被 p 个参数 $\hat{\beta}$ 刻画。由最小二乘估计的性质知，由（6）得到的参数的估计是无偏估计，而且当数据满足线性模型的诸假定时，这个估计是最优的。

图 1 是一个用线性模型分类的例子[4]。在二维平面上随机产生两组数据，这两组数据均服从正态分布，均值分别为 0 和 1，每组数据包含 100 个点。图 1 表示这些点的散点图，输入是一个二维变量 (X_1,X_2)，输出类变量 G 有两个取值 0 或 1，在图 1 中分别用 blue 或 red 表示，每个类都有 100 个点。用线性回归拟合这些数据，由公式（2）得到的拟合值 $\hat{Y}$ 根据以下规则转换到拟合类变量 $\hat{G}$：

$$\hat{G}=\begin{cases}red, & if\ \hat{Y}>0.5;\\ blue, & if\ \hat{Y}\le 0.5\end{cases}$$

二维平面中的点 $x\in R^2$ 集合由判定边界 $\{x:x^T\beta=0.5\}$ 分开，此时，边界是线性的。判定边界上方表示被分类为 red 的部分，判定边界下方表示被分类为 blue 的部分。

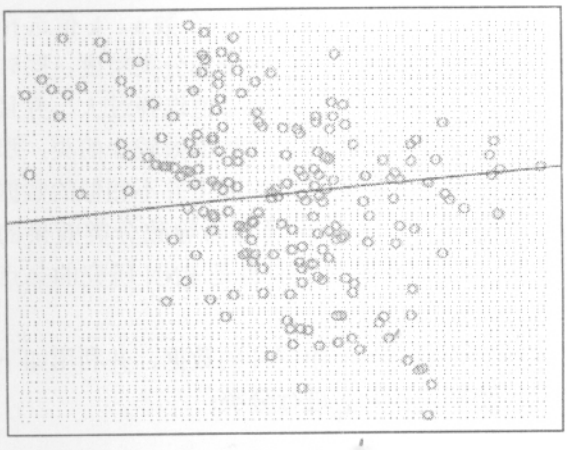


图 1　一个二维空间上的分类例子。类用二元变量编码（blue=0，red=1）。直线是 $\{x:x^T\beta=0.5\}$ 定义的判定边界。直线上方表示被分类为 red 的部分，直线下方表示被分类为 blue 的部分。

由图 1 显示的结果可以看出每个类都有一部分数据点被误分，总的来讲，在训练集上的误分率达到了 14%，这是一个比较大的错误率。故此，对于基于最小二乘方法得到的线性分类器来说，虽然该分类器简单、计算量小、估计结果比较稳定，方差小，但是它的灵活性不够，误分率也比较大。

2　k-最近邻 (KNN) 分类器

另一种常见的分类器是 K-最近邻分类器。K-最近邻分类器又简记为 KNN，最初由 Cover 和 Hart 于 1968 年提出的，是一个理论上比较成熟的方法[5]。该方法的思路非常简单直观：如果一个样本在特征空间中的 k 个最相似 (即特征空间中最邻近) 的样本中的大多数属于某一个类别，则该样本也属于这个类别。具体来说，假设 m 维欧式空间 R^m 中的两个点 a 和 b，它们之间的欧式距离定义如下：

$$d(a,b)=\sqrt{\sum_{i=1}^{m}(a_i-b_i)^2}$$

对于训练集中的每一个输入 x_i，我们可以按如下公式得到它的拟合 $\hat{Y}$：

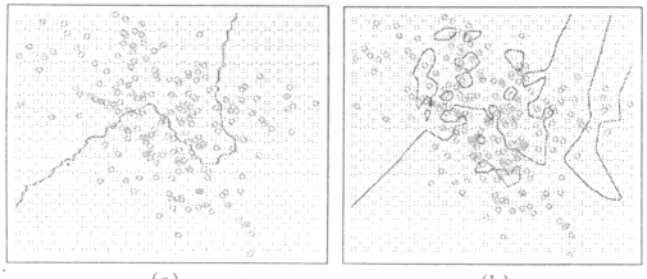
$$\hat{Y}(x_i)=\frac{1}{k}\sum_{x_i\in N_k(x_i)}y_i \qquad (7)$$

其中，$N_k(x)$ 是 x 的 k-邻域，由训练样本中最邻近 x 的 k 个点 x_i 定义。(7) 的实质是找出输入空间与 x 最近邻的 k 个观测值 x_i，并对它们的响应取平均值。如果在观测 x 邻域中某一类明显占优势，则 $\hat{Y}(x_i)$ 的值更接近该类别的取值，故观测样本也更可能属于该类。

与最小二乘方法必须拟合 p 个参数不同，KNN 算法只有一个参数，即邻居的个数 k。k 值选择过小，得到的近邻数过少，会降低分类精度，同时也会放大噪声数据的干扰；而 k 值选择过大，并且待分类样本属于训练集中包含数据数较少的类，那么在选择 k 个近邻的时候，实际上并不相似的数据亦被包含进来，造成噪声增加而导致分类效果的降低。如何选取合适的 k 值也是 KNN 研究的热点问题之一[6,7]。

我们使用与图 1 中同样的数据，并利用如下规则对任意新的输入 x_0 进行分类：

$$\hat{G}=\begin{cases}red, & if\ \hat{Y}(x_0)>0.5;\\ blue, & if\ \hat{Y}(x_0)\le 0.5\end{cases}$$



(a) (b)

图 2 与图 1 相同的二维分类例子。(a) 15-最近邻 (b) 1-最近邻分类预测。

其中 $\hat{Y}(x_0)$ 由 (7) 定义。图 2 是 15-最近邻 (a) 和 1-最近邻 (b) 的分类结果。相对于图 1，(a) 中的边界曲线光滑度变差，稳定性也变差，数据的误分率变小。在 (b) 中，边界更加不光滑，但是没有一个数据被误分。

3 两种分类算法的比较

对于前面介绍的两种分类方法，我们模拟产生图 1 中的数据，在 200 个数据点上进行训练，在 10000 个数据点上进行测试，得到的统计结果如表 1 所示。

表 1 两种分类方法的统计结果比较

方法	预测误差	
	训练集	测试集
线性回归	0.14	0.185
Knn(15)	0.11	0.135
Knn(1)	0.0	0.165

由表 1 得结果可见，用线性回归进行分类，在训练集和测试集上的预测误差都比较大。利用 KNN (15) 方法进行分类，在训练集和测试集上的预测误差都小于线性回归得到的预测误差，即该分类的精度优于线性回归。利用 KNN (1) 进行分类，在训练集上没有预测误差，但是在测试集上的预测误差却超过了 KNN (15)。这理由是显然的，因为训练集上每个数据的邻域就是它本身，当然没有预测误差。但是对于测试集而言，由于存在噪声，利用训练集上的结果去对测试集进行分类肯定存在预测误差。

最小二乘法得到的线性判定边界非常光滑，并且对于拟合显然是稳定的。由于对于全部数据只需要拟合一次回归直线，故该算法的计算量比较小。但是该方法过分依赖于线性判定边界是合适的假定，它具有低方差和潜在的高偏倚。

另一方面，k-最近邻过程不依赖于对基础数据的任何严格假定，并能适合任何情况。然而，判定边界的任何特定子部分都依赖于少数点和它们的特定位置，并因而是摆动和不稳定的——具有高方差和低偏倚。另外，由于 KNN 方法主要靠周围有限的邻近的样本，而不是靠判别类域的方法来确定所属类别的，因此对于类域的交叉或重叠较多的待分样本集来说，KNN 方法较其他方法更为适合。再者，该方法的计算量较大，因为对每一个待分类的数据都要计算它到全体已知样本的距离，才能求得它的 K 个最近邻点。现在已经提出很多的优化算法来降低 KNN 算法的计算复杂度，提高分类的效率。

综上所述，线性回归分类器和 KNN 分类器的比较如表 2 所示。

表 2　线性回归分类器和 KNN 分类器的比较

方法	优点	缺点
线性回归	1. 计算简单 2. 拟合具有低方差	1. 边界不够灵活 2. 对于训练集误判率比较大 3. 依赖于线性边界合适的假定
KNN	1. 边界比较灵活 2. 对于训练集误判率比较小 3. 拟合偏差比较小 4. 不依赖对基础数据的任何假定	1. 计算量比较大 2. 拟合具有高方差 3. 对于测试集误判率依赖于 k 的选取

4　结论

本文从理论上和数值模拟上比较了两种简单的分类器：线性回归分类器和 KNN 分类器。线性分类器计算简便，拟合具有低方差，如果需要分类的数据来自不同的类别，且类别之间相互重叠的区域比较小时，线性分类器是一个合适的选择。相对于线性分类器，KNN 分类器分类更灵活，拟合偏差比较小，由于计算量比较大，该算法更适合于类别界限不是很明显，数据之间交叉或重叠比较多的数据集。

【参考文献】

[1] 梁晓音.机器学习在数据挖掘中的应用[J].广西质量监督导报，2008，(11)：38–40.

[2] Kutner,Nachtsheim.应用线性回归[M].北京：高等教育出版社，2005.

[3] 陈明，何书萍，李凡长. 一种李群机器学习线性分类算法研究[J].微电子学与计算机，2009，(10)：170–173.

[4] 黑斯蒂.统计学习基础 数据挖掘、推理与预测[M].北京：电子工业出版社，2004.

[5] 潘丽芳.基于簇的 KNN 分类算法研究[J]. 计算机工程与设计，2009，(18)：4260–4262.

[6] 闭小梅，闭瑞华. KNN 算法综述[J]. 科技创新导报，2009，(14)：31.

[7] 张著英,黄玉龙,王翰虎.一个高效的 KNN 分类算法[J]. 计算机科学，2008，(3)：170–172.