基于K-means的朴素贝叶斯分类算法的研究

张亚萍[1],胡学钢[2]
(1.淮北煤炭师范学院 物理系,安徽 淮北 235000;
2.合肥工业大学 计算机与信息学院,安徽 合肥 230009)

摘 要:将K-means算法引入到朴素贝叶斯分类研究中,提出一种基于K-means的朴素贝叶斯分类算法。首先用K-means算法对原始数据集中的完整数据子集进行聚类,计算缺失数据子集中的每条记录与 k 个簇重心之间的相似度,把记录赋给距离最近的一个簇,并用该簇相应的属性均值来填充记录的缺失值,然后用朴素贝叶斯分类算法对处理后的数据集进行分类。实验结果表明,与朴素贝叶斯相比,基于K-means思想的朴素贝叶斯算法具有较高的分类准确率。

关键词:朴素贝叶斯分类;k-means算法;缺失数据
中图分类号:TP301.6 **文献标识码**:A **文章编号**:1673-629X(2007)11-0033-03

Research of Naive Bayesian Classification Based on K-means Method

ZHANG Ya-ping[1],HU Xue-gang[2]
(1.Department of Physics, Huaibei Coal Industry Teachers' College,Huaibei 235000,China;
2.School of Computer & Information,Hefei University of Technology,Hefei 230009,China)

Abstract:A naive Bayesian classification based on K-means method by introducing clustering algorithm into naive Bayesian classification. The similarity between every recorder in absent data subsets and the centers k cluster is calculated by clustering complete data subsets of initial data by K-means algorithm,then the recorder is set to the nearest cluster and the absent value of the record is filled by the corresponding attribute of the cluster,finally,the handled data set is clustered by naive Bayesian classification algorithm. The experiments show that naive Bayesian classification based on K-means method has the higher precise of clustering comparing with naive Bayesian classification.

Key words:naive Bayesian classification;K-means algorithms;missing data

0 引 言

为了提高朴素贝叶斯的分类准确度,已有许多相关文献在放宽条件独立性的限制方面作了一些改进。实际上影响朴素贝叶斯分类准确度的不仅仅是属性之间的相关性,还有数据的完备性,所以缺失数据[1]的处理效果对朴素贝叶斯分类准确度有很大影响。基于这个原因把聚类中的K-means算法引入朴素贝叶斯分类中,来提高朴素贝叶斯分类的准确度。

1 朴素贝叶斯的分类过程

① 给定一个没有类标号的数据样本 $\boldsymbol{X}$,用 n 维特征向量 $\boldsymbol{X}=\{x_1,x_2,\cdots,x_n\}$ 表示,分别描述样本 $\boldsymbol{X}$ 在 n 个属性 $\{A_1,A_2,\cdots,A_n\}$ 上的属性值。假定有 m 个类 $\{C_1,C_2,\cdots,C_m\}$,那么,将样本 $\boldsymbol{X}$ 分配给类 C_i 条件就是:

$$P(C_i/\boldsymbol{X})>P(C_j/\boldsymbol{X}) \quad (1\leqslant j\leqslant m, j\neq i)$$

即假定样本为类 C_i 的概率大于假定为其它类的概率。根据贝叶斯定理[2]:

$$P(C_i/\boldsymbol{X})=\frac{P(\boldsymbol{X}/C_i)P(C_i)}{P(\boldsymbol{X})}$$

其中,$P(\boldsymbol{X})$ 指任意一个数据对象符合样本 $\boldsymbol{X}$ 的概率,对于所有类来说,它为常数。由公式可看出,只需要 $P(\boldsymbol{X}/C_i)P(C_i)$ 最大即可。$P(C_i)$ 为任意一个数据对象是类 C_i 的概率,可以用 $P(C_i)=s_i/s$(其中,s_i 是类 C_i 中训练样本数,s 是训练样本总数)计算。

② 给定样本的类标号,假定各属性值相互条件独立(类条件独立),这样 $P(\boldsymbol{X}/C_i)$ 的计算可用公式:

$$P(\boldsymbol{X}/C_i)=\prod_{k=1}^{n}P(x_k/C_i)$$

概率 $P(x_k/C_i)$ 可以由训练样本估算:

如果 A_k 是离散属性,则 $P(x_k/C_i)=s_{ik}/s_i$,其中

收稿日期:2007-01-29
基金项目:安徽省自然科学基金资助项目(050420207)
作者简介:张亚萍(1978-),女,安徽人,讲师,研究方向为人工智能与数据挖掘;胡学钢,博士,教授,研究方向为人工智能与数据挖掘。

s_{ik} 是属性 A_k 上值为 x_k 的类 C_i 的训练样本数，s_i 是 C_i 中的训练样本数。

如果 A_k 是连续值属性，那么假设属性服从高斯分布。因此就有：

$$P(x_k/C_i) = g(x_k, \mu_{c_i}, \sigma_{c_i}) = \frac{1}{\sqrt{2\pi}\sigma_{c_i}} e^{\frac{(x-\mu_{c_i})^2}{2\sigma_i}}$$

其中：$g(x_k, \mu_{c_i}, \sigma_{c_i})$ 为属性 A_k 的高斯规范密度函数；μ_{c_i} 和 σ_{c_i} 为训练样本中类别为 C_i 的属性 A_k 的均值和方差。

③ 为未知样本 $\boldsymbol{X}$ 分类，对每个类 C_i 计算 $P(\boldsymbol{X}/C_i)P(C_i)$。样本 $\boldsymbol{X}$ 被指派到类 C_i，当且仅当 $P(\boldsymbol{X}/C_i)P(C_i) > P(\boldsymbol{X}/C_j)P(C_j), 1 \leqslant j \leqslant m, j \neq i$。

朴素贝叶斯分类假定类条件独立，简化了计算。当假定成立时，与其它分类算法相比，朴素贝叶斯分类是最精确的[3]，同时在处理大规模数据库时表现出了较高的分类准确性和运算性能，它还可为其它分类算法提供理论判断。但是缺失数据的存在对分类的准确度有很大的影响，基于这些原因文中提出一种朴素贝叶斯分类改进算法。

2 基于 K－means 的朴素贝叶斯分类算法的基本步骤

算法主要步骤如下：

输入：$D = \{X_1, X_2, \cdots, X_n\}$，其中，$X_1, X_2, \cdots, X_n$ 为原始属性集，$C = \{C_1, C_2, \cdots, C_m\}$ 为类别属性，簇的数目为 K

输出：样本 $\boldsymbol{X}$ 的类别号

步骤一：把数据集 D 分为两个数据子集 D_c 和 D_i。D_c 中的记录全部为完整记录，没有任何属性含有缺失值；D_i 中的记录为缺损记录，即属性中含有一个及以上的缺失值。

步骤二：对数据子集 D_c 使用 K－means 算法：任意选择 k 个对象作为初始的簇重心；对数据集中的各个对象，计算其与重心的距离，将每个对象赋给最近的簇；然后更新簇的平均值，即计算每个簇中对象的平均值，直到目标函数达到最小[4]。

步骤三：从数据子集 D_i 中顺序取出记录，计算该记录与 D_c 的 k 个类中任一类的相异度，选出最小的相异度，把该记录标记为 $K_i(i = 1,2,\cdots,k)$ 类；直到数据子集 D_i 为空。

步骤四：根据 D_i 中记录被分配的类，对记录的缺失值进行如下处理：

$$\text{缺失值} = \begin{cases} \dfrac{\sum_{i=1}^{n} A_i}{n} & \text{缺失值为数值型数据} \\ A_i & \text{缺失值为离散型数据} \\ & \text{取出频率最多的 } A_i \end{cases}$$

其中 n 为属于该类的记录总数，A_i 为相应的属性值。

步骤五：计算后验条件概率 $P(C_i/\boldsymbol{X})$；

```
for (j = 1; j < m; j ++)
  { 计算 P(C_j);
    while (1 < i < l)
    {for (k = l; k <= S_i; k ++)
       计算 p(x_k/C_j) = S_ik/S_i;
    i ++;}}
```

$$\text{计算 } P(C_i/\boldsymbol{X}) = P(C_i)\prod_{k=1}^{n} P(x_k/C_i)$$

步骤六：分配类标号：

```
for (i = 1; i < m; i ++)
  {if Max (P(C_i/X))
  then 样本 X 类标号为 C_i;}
```

步骤七：结束，输出类标号 C_i。

3 基于 K－means 的朴素贝叶斯分类算法的应用

表 1 是某高校某届学生的基本信息。由于种种原因一些学生的信息有缺失，如果直接用朴素贝叶斯分类算法预测学生毕业后考研/就业情况，将会影响预测的准确度。现用基于 K－means 的朴素贝叶斯分类算法提高预测的准确度。

表 1 某校某届学生的基本信息

学号	高数	普物	大学英语	工程制图	电路分析	C语言	线性电子线路	数字电路	…	就业/考研
1	及格	中	中	中	及格	及格	及格	及格	…	E
2	及格	及格	及格	中	中	及格	及格	及格	…	E
3	良	中	及格	优	良	及格	良	优	…	E
4	优	良	中	良	优	良	优	优	…	S
5	及格	及格	及格	及格	及格	及格	及格	及格	…	E
6	良	中	及格	中	良	中	优	优	…	S
7	良	中	及格	中	优	及格	优	中	…	E
8	良	良	中	及格	良	中	中	优	…	E
9	优	良	中	良	优	中	优	优	…	E
10	及格	良	中	中	及格	及格		中	…	E
11	优	良	中	良	良	良	良	优	…	S
12	良	良		优	优	良	良	良	…	E
13	及格	及格	及格	及格	良	及格	中	良	…	E
14	中	及格	及格	良	中	良	中	良	…	E
15	良	中	良	良	优	中	良	良	…	S
16	中	中	中	良	优	良	良	良	…	E
17	中	良		良	良	及格	良	良	…	E
18	优	优	中	良	优	良	优	优	…	S
19	及格	中	中	良	优	良	良	优	…	S
20	良	良	良	良	良	中	中	良	…	E
…	…	…	…	…	…	…	…	…	…	…

(1) 每个数据样本均是由一个 14 维特征向量 $\boldsymbol{X}=\{x_1,x_2,x_3,\cdots,x_{14}\}$ 表示,分别描述对 14 个属性 $A_1,A_2,A_3,\cdots,A_{14}$ 的 14 个度量,即分别描述对数据样本 14 个属性:高等数学、普通物理、大学英语、……、信息论 14 个度量。

(2) 用聚类中的 K－means 算法对完整数据子集进行聚类,然后计算缺失子集的元素与 K 个类的相异度,根据计算的结果填充记录的缺失值。

(3) 共有 2 个不同类别 C_1,C_2。C_1 值(就业)为"E",C_2 值(考研)为"S"。$p(C_i)=s_i/s$,其中 s_i 为类 C 中的训练样本数,s 为训练样本总数。

$Pe=p$(就业/考研="E")=(就业/总人数)

$Ps=p$(就业/考研="S")=(考研/总人数)

(4) 假定属性值相互条件独立,给定样本的类标号,则 $P(\boldsymbol{X}/C_i)=\prod_{k=1}^{n}P(x_k/C_i)$。概率 $P(x_k/C_i)$ 可以由训练样本估值,其中:A_k 是符号量,则 $P(x_k/C_i)=s_{il}/s_i$,其中 s_{il} 为训练样本中类别为 C_i 且属性 A_k 取 v_k 值的样本数,s_i 是训练样本中类别为 C_i 中的样本数。

表 2 是用朴素贝叶斯与基于 K－means 思想的朴素贝叶斯分类准确度的比较。实验结果表明:基于 K－means 思想的朴素贝叶斯分类在实验数据集上取得较好的分类性能,其分类性能明显优于朴素贝叶斯分类。在 k 取值不同时,分类准确度也有所不同,这说明选取合适的 k 值,有助于分类的准确度的提高。

表 2　朴素贝叶斯与基于 K－means 朴素贝叶斯分类准确度比较

朴素贝叶斯分类	基于 K－means 思想的朴素贝叶斯分类		
	$K=4$	$K=5$	$K=6$
76.22%	77.32%	78.14%	76.37%

4　总　结

将聚类中的 K－means 算法与朴素贝叶斯分类方法相结合,提出一种基于 K－means 的朴素贝叶斯分类算法。实验表明,此算法分类的准确率优于朴素贝叶斯分类算法。但此改进算法只与传统的朴素贝叶斯算法作了比较,和其它一些分类技术相比,有没有优越性,性能如何,这也是下一步将要考虑的问题[5]。

参考文献:

[1] Han Jiawei, Kamber M. 数据挖掘概念与技术[M]. 第 2 版. 范　明,孟小峰,等译. 北京:机械工业出版社,2001.

[2] 姜卯生. 数据挖掘中基于贝叶斯技术的分类问题的研究[D]. 合肥:合肥工业大学,2004.

[4] Semi K L. Naive Bayesian Classifiers[C]//In: Proceedings of European Conference on Artificial Intelligence. Porto, Portugal: Springer Verlag, 1991:206－219.

[3] 黄　捷. 数据挖掘技术中的贝叶斯分类算法研究[D]. 广州:华南理工大学,2001.

[5] 郭亚光. 基于粗糙集合和朴素贝叶斯模型的分类问题研究[D]. 合肥:合肥工业大学,2005.

(上接第 32 页)

况车牌图片进行定位实验。如图 7 所示,大图的左部

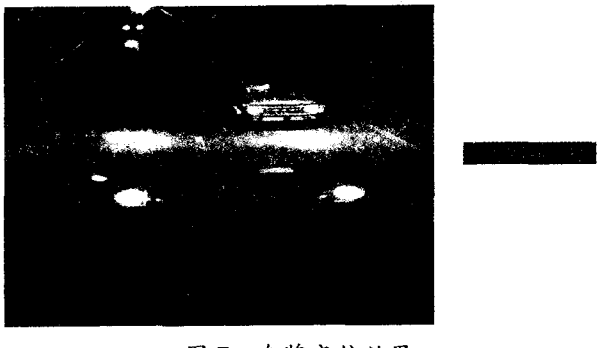

图 7　车牌定位效果

是全景,右部是抓拍的近景(车牌定位、识别即在此图中进行),小图是定位结果,可以看出,定位准确,不受灯光、营运证等干扰。共计两千四百五十张图片,除去本来不含车牌的,定位成功率约为 99%。在 P4 1.8G CPU、512MB 内存的台式计算机 Matlab 环境下运行,平均定位时间小于 0.15 秒/帧。这个结果相对于纯粹基于纹理的自适应方法和纯粹基于颜色的形态学方法所能达到的结果要高至少十个百分点,而所花费的时间却是后两种方法的五分之一左右。表明本算法的定位准确率、鲁棒性、处理速度等都达到了实用化程度。

参考文献:

[1] 李　波,曾致远,周建中,等. 车牌识别系统研究与实现[J]. 计算机技术与发展,2006,16(6):10－14.

[2] 贾小军,王晓燕,喻擎苍. 一种基于纹理模式的汽车牌照定位方法[J]. 微计算机信息,2006,22(10):240－242.

[3] 张　引,潘云鹤. 彩色汽车图像牌照定位新方法[J]. 中国图象图形学报,2001,6(4):374－377.

[4] 郭大波. 彩色汽车图像车牌定位技术分析[J]. 山西大学学报:自然科学版,2005,28(1):40－43.

[5] Cao Guangzhi, Chen Jianqian, Jiang Jingping. An adaptive approach to vehicle license plate localization[C/OL]. Industrial Electronics Society, 2003. IECON '03. The 29th Annual Conference of the IEEE. Vol. 2: 1786－1791. http://ieeexplore.ieee.org/xpls/abs-all.jsp?arnumber=1280330.