云南大学学报（自然科学版），2011，33（5）：517～520　　CN 53－1045/N　ISSN 0258－7971
Journal of Yunnan University　　http://www.yndxxb.ynu.edu.cn

基于条件互信息下聚类的朴素贝叶斯分类算法*[1]

彭兴媛，刘琼荪，王立威
（重庆大学 数学与统计学院，重庆　401331）

摘要：采用条件互信息来度量任意2个条件属性之间的关联程度，采用互信息度量各条件属性与类属性间的关联程度，以此作为将各条件属性进行聚类的准则，提出一种新的将条件属性进行聚类的分组技术. 同时，结合朴素贝叶斯分类算法，构造了改进的朴素贝叶斯分类模型. 通过仿真实验表明该文提出的算法具有较好的分类性能.

关键词：关联程度；聚类算法；条件互信息；互信息
中图分类号：TP 301.6　　**文献标识码**：A　　**文章编号**：0258－7971(2011)05－0517－04

分类问题是数据挖掘和机器学习领域的核心问题，目前已有许多分类算法，如决策树、神经网络、支持向量机、贝叶斯方法等，其中贝叶斯分类器由于能综合先验信息和数据样本信息，成为当前机器学习和数据挖掘的研究热点之一.

朴素贝叶斯分类方法[1]因其具有坚实的数学理论基础使得其性能可与决策树、神经网络等方法相媲美，甚至在某些领域中能够表现出更加优越的性能，因此是目前公认的一种简单有效的分类方法. 然而，朴素贝叶斯分类方法中的“属性间的独立性假设”在实际应用问题中难以满足，容易导致分类效果不佳. 为了弱化独立性假设的条件，许多学者做了大量的研究. 如 Kononenko[2] 提出一种采用穷尽搜索的属性分组技术，使得属性间的独立性条件弱化为属性组之间的独立性. 但运用到现实问题中时，属性能够被完全分成独立的子集合也只是少数情况. 文献[3]通过计算出无缺损值属性间的相关系数作为属性间的相关程度进行分析聚类后，提出一种改进的贝叶斯算法，用于对缺损数据的修补，文献[4]在现有对空间分类算法研究的基础上，提出一个基于表示对象之间的邻接关系的邻接图及朴素贝叶斯分类法的新的空间分类算法.

本文在文献[3]的聚类分析思想的基础上，不仅以信息论的互信息作为度量条件属性和类属性的重要度，还考虑了用条件互信息来度量条件属性间的关联程度，作为聚类的参考值，并结合 SNBC 分类思想[2]提出一种改进了的朴素贝叶斯分类模型，INB 代表本文分类模型. 实验表明，此方法可以有效地提高分类效果.

1　准备知识

1.1　朴素贝叶斯分类模型　假设 A 为数据集中有 n 个条件属性的集合 $A=\{A_1,A_2,\cdots,A_n\}$，且数据集中类属性 C 为包含有 m 个不同取值的集合，记为 $C_1,C_2,\cdots,C_m$. 朴素贝叶斯分类算法就是将数据集中的一个待分类样本 $X=\{a_1,a_2,\cdots,a_n\}$（其中 a_1 是条件属性 A_i 的取值）分配给某个类 $C_i(1\leqslant i\leqslant m)$，当且仅当 $P(C_i|X)>P(C_j|X)(1\leqslant i,j\leqslant m, j\neq i)$. 根据贝叶斯定理，有

$$P(C_i|X)=\frac{P(C_i)P(X|C_i)}{P(X)},$$

因 $P(X)$ 对于所有的类标签均为常数，故有

$$P(C_i|X)=\frac{P(C_i)P(X|C_i)}{P(X)}\propto P(C_i)P(X|C_i). \tag{1}$$

* **收稿日期**：2011－03－31
基金项目：中央高校基本科研业务费资助(CDJXS11 10 00 50).
作者简介：彭兴媛(1985－)，女，硕士，主要研究方向：数据分析和数据挖掘.
通讯作者：刘琼荪(1956－)，女，教授，主要研究方向：智能计算、数据挖掘和应用统计.

由条件属性相互独立的假设,有 $P(C_i|X) \propto P(C_i)\prod_{k=1}^{n}P(\alpha_k|C_i)$,其中 $P(C_i) = \frac{S_i}{S}$,S_i 是类 C_i 中的训练样本数,S 是训练样本总数. 则朴素贝叶斯分类模型表示为:

$$C_{nb}(X) = \underset{C_i \in C}{\arg\max} P(C_i)\prod_{k=1}^{n}P(\alpha_k|C_i). \quad (2)$$

1.2 SNBC 分类模型 由 Kononenko 提出的 SNBC 分类模型是将属性中依赖关系较大的属性聚类为一个子集,且子集与子集间是相互独立的,这就将属性间的独立性弱化到了属性子集之间的独立性. 基于某些标准构成属性子集组合 $B = \{B_1, B_2, \cdots, B_p\}(1 \leqslant i \leqslant p \leqslant n)$,其中 $B_i = \{A_{i_1}, A_{i_2}, \cdots, A_{i_k}$ 是原属性集 $\{A_1, A_2, \cdots, A_n\}$ 的一个属性子集. 满足 $B_1 \cup B_2 \cup \cdots \cup B_p = \{A_1, A_2, \cdots, A_n\}$ 和 $B_i \cap B_j = \varnothing$,当 $i \neq j, 1 \leqslant i, j \leqslant p$ 时. 假设类属性为 C,且 B_i 和 B_j 只是相对于类属性 C 而言是条件独立的,即有:$P(B_i, B_j|C) = P(B_i|C)P(B_j|C)(1 \leqslant i, j \leqslant p)$,则 SNBC 分类模型可描述为:

$$C_{xnbc}(X) = \underset{C_i \in C}{\arg\max} P(C_i)\prod_{j=1}^{P}P(b_j|C_i). \quad (3)$$

这里的 b_j 是属性子集 B_j 中包含的各个条件属性 $A_{j1}, A_{j2}, \cdots, A_{jk}$ 分别取值 $a_{j1}, a_{j2}, \cdots, a_{jk}$ 时的数值向量,即 $b_j = \{a_{j1}, a_{j2}, \cdots, a_{jk}\}$.

2 相关概念

定义 1 (条件互信息)[5] 设条件属性 $A_i = (a_{j1}, \cdots, a_{in})$,$A_j(a_{j1}, \cdots, a_{jm})$ 其中 a_{ik}, a_{jv} 分别为属性 A_i 和 A_j 具体的属性值,类属性 $C = (C_1, \cdots, C_m)$,则定义属性 A_i 和 A_j 的条件互信息函数为:

$$I(A_i;A_j) = \sum_{a_{ik},a_{jv},C_h} P(a_{ik}, a_{jv}, a_h) \cdot \log_2 \frac{P(a_{ik}, a_{jv}|C_h)}{P(a_{ik}|C_h)P(a_{jv}|C_h)} = \sum_{a_{ik},a_{jv},C_h} P(a_{ik}, a_{jv}, a_h) \cdot \log_2 \frac{P(a_{ik}, a_{jv}, C_h)P(C_h)}{P(a_{ik}, C_h)P(a_{jv}, C_h)}. \quad (4)$$

其中 $P(a_{ik}, a_{jv}, a_h)$ 表示属性值 a_{ik}, a_{jv}, a_h 同时出现的概率.

本文采用条件互信息来衡量条件属性间的关联程度. 条件互信息 $I(A_i;A_j|C)$ 具有对称性,因此当类属性 C 给定后,可从条件属性 A_j 处获得条件属性 A_i 的信息量,或者从条件属性 A_i 处获得条件属性 A_j 的信息量,因此 $I(A_i;A_j|C)$ 反映了在给定类属性 C 的情况下条件属性 A_i 与 A_j 之间的相互关联程度的大小,条件互信息值越大,关联程度越大. 特别地,当条件属性 A_i 和 A_j 在类属性 C 下是条件独立时,有 $I(A_i;A_j|C) = 0$.

定义 2 (互信息)[5] 条件属性 A_i 与类属性 C 的互信息定义为:

$$I(A_i;C) = \sum_{a_{ik},C_h} P(a_{ik}, C_h)\log_2 \frac{P(a_{ik}|C_h)}{P(a_{ik})} \sum_{a_{ik},C_h} P(a_{ik}, C_h)\log_2 \frac{P(a_{ik}, C_h)}{P(a_{ik})P(C_h)}. \quad (5)$$

互信息表示从类属性 C 所获得的关于条件属性 A_i 的平均信息量. $I(A_i;C)$ 的值越大,说明条件属性 A_i 对于类属性 C 就越重要[6]. 特别地,当 A_i 与 C 独立时,则 $I(A_i;C) = 0$.

3 聚类算法步骤

本文以条件互信息和互信息分别度量条件属性间和条件属性与类属性间的关联程度大小作为聚类的参考值. 如当 $I(A_j;A_i|C) > I(A_j;A_k|C)$ 时,说明当类 C 给定后 A_j 对 A_i 的依赖程度大于 A_j 对 A_k 的依赖程度,如当 $I(A_i;A_j|C) > I(A_j;C)$ 时,说明 A_j 对 A_i 的依赖程度大于 A_j 对类属性 C 的重要度. 因此判断 A_j 与 A_i 聚为一属性子集的可能性较大.

步骤 1 计算所有条件属性间的条件互信息值和所有条件属性与类属性间的互信息值;

步骤 2 找出具有最大值的 $I(A_i;A_j|C)$(如有多个相同最大值则任取 1 个)的条件属性 A_i 和 A_j,此时 $A_i \in A, A_j \in A$,如果 $I(A_i;A_j|C) > I(A_i;C)$ 且 $I(A_i;A_j|C) > I(A_j;C)$,则 A_i 和 A_j 聚为一个属性子集 B_1,否则 A_i 和 A_j 各自聚为一个属性子集 B_i, B_j.

因步骤 2 产生了属性子集,此时 $I(A_p;A_q|C)$ 的取值方式因 A_p, A_q 可能出现如下 3 种情况:

① 当 $A_p, A_q \in A$ 且 $A_p, A_q \notin B_j, \forall j, I(A_i;A_j|C)$ 由公式(4)计算;

② 当 $A_p \in B_l, A_q \in B_m$ 时,$I(A_p;A_q|C) = \min\{I(A_r;A_k|C), \forall A_r \in B_e, \forall A_k \in B_m\}$;

③ 当 $A_p \in B_l, A_q \in A$,且 $A_q \notin B_j, \forall j$ 时,$I(A_p;A_q|C) = \min\{I(A_r;A_q|C), \forall A_r \in B_l\}$;

步骤 3　求出$\max\limits_{A_p,A_q} I(A_p;A_q \mid C)$,此时$\max\limits_{A_p,A_q} I(A_p;A_q \mid C)$可能出现以下 3 种情况:

(1) 在情况 ① 下,则返回到步骤 2;

(2) 在情况 ② 下,则停止聚类,且剩下未被聚类的条件属性各自聚为一个属性子集;

(3) 在情况 ③ 下,若 $I(A_p;A_q \mid C) > I(A_q;C)$ 且 $I(A_p;A_q \mid C) > I(A_h;C)$,$(\forall A_h \in B_l)$ 则将 A_q 聚到属性子集 B_l 中去,否则 A_q 自聚成另一个属性子集 B_q.

4　INB 分类算法实现步骤

步骤 1　数据预处理:将训练样本和待分类样本进行补齐和离散化.

步骤 2　分类模型构造:

(1) 计算条件属性间的条件互信息和条件属性与类属性间的互信息;

(2) 采用 3 的聚类算法步骤得到最终的属性子集 $B_1,B_2,\cdots,B_p \cdot (1 \leqslant p \leqslant n)$;

(3) 扫描所有的训练样本,统计训练集中类标签为 C_i 时属性子集 B_j 取值 b_j 的实例数 $S(b_j;C_i)$;并计算所有的先验概率 $P(b_j \mid C_i)$,有 $P(b_j \mid C_i) = \dfrac{S(b_j;C_i)}{S_i}$;以及 $P(C_i)$,即取值为类标签 C_i 的概率;

(4) 生成朴素贝叶斯概率表.

步骤 3　根据公式(3),计算 $P(C_i \mid X)$,当 $P(C_i \mid X) > P(C_j \mid X)$,$\forall i \neq j$,判断 $X \in C_i$.

5　仿真实验

为了验证算法的可行性和有效性,对本文算法进行了仿真实验,并与单纯的朴素贝叶斯算法进行比较. 实验数据来源于 UCI[7] 数据库. 本算法主要处理的是离散数据,因此在进行实验前需要对所选数据进行离散化的预处理,即对连续数据实行分段离散化处理. 在相同的试验环境下,利用 MATLAB 编程分别实现了 NB 算法(朴素贝叶斯算法)和本文提出的 INB 算法. 实验前,首先对数据进行随机扰动,并对数据集采用 10 次交叉验证,每个数据集轮流进行 10 次,取其平均值作为实验测试结果,其结果见表 1.

6　结论及展望

本文以条件互信息度量条件属性间的关联程度,以互信息度量各条件属性与类属性间的关联程度,提出了一种新的聚类方法即属性间的分组技术,并用于朴素贝叶斯分类模型,克服了朴素贝叶斯分类模型的独立性假设条件的缺陷. 通过仿真实验,本文提出的算法与朴素贝叶斯分类算法进行比较,取得了较好的分类效果. 但 INB 算法仅局限于离散化的数据,因此采用不同的离散方法得到的分类效果也会不同.

表 1　实验测试结果

Tab. 1 The results of determination

数据集	实例数	类数	属性数	NB 正确率/%	INB 正确率/%
Car Evlauation	1 728	4	6	85.735 1	90.915 0
Teaching Assistant Evaluation	151	3	5	51.982 1	68.919 6
Tic – Tac – Toe	958	2	9	69.651 9	73.714 3
Postoperative Patient	90	3	8	62.469 1	65.676 6
Zoo	101	7	16	90.186 7	95.271 6
Haberman's Survival	306	2	3	70.861 4	73.126 6

参考文献:

[1] HAN J W, KAMBER M. Data mining concepts and techniques[M]. New York: Morgan Kaufmann Publishers, 2001.

[2] KONONENKO I. Semi – naive bayesian classifier[C]// Proceedings of the 6th European Working Session on Learning. New York: Springer – Verlag, 1991: 206-219.

[3] 余瑞康. 聚类思想在贝叶斯算法中的应用[J]. 计算机工程与应用, 2006, 28: 159-163.

[4] 赵秦怡, 王丽珍, 周丽华. 一种基于朴素贝叶斯分类

法的空间分类算法[J]. 云南大学学报:自然科学版,2004,26(4):297-300.
[5] 陈前斌,蒋青,于秀兰. 信息论基础[M]. 北京:高等教育出版社,2008.
[6] 王国胤,于洪,杨大春. 基于条件信息熵的决策表约简[J]. 计算机学报,2002,25(7):759-766.
[7] BLAKE C L, MERZ C J. UCI Repository of Machine Learning Databases [EB/OL]. 1998 - 12 - 30. http://www.ics.uci.edu/mlearn/MLRepository.html.

Naive Bayesian classification algorithm based on clustering with conditional mutual information

PENG Xing-yuan, LIU Qiong-sun, WANG Li-wei
(College of Mathematics and Statistics, Chongqing University, Chongqing 401331, China)

Abstract: In this paper, the correlation intensity of two arbitrary conditional attributes was measured by conditional mutual information, and the correlation intensity between every conditional attribute and classification attribute was measured by mutual information. On that criterion to cluster the conditional attributes, a new grouping method to cluster the conditional attributes was proposed. Simultaneously, combined with naive bayes classification algorithm, an improved naive bayes classification model was constructed. Simulation results showed the efficiency of this method is preferable.

Key words: correlation intensity; clustering algorithm; conditional mutual information; mutual information

* *

(上接第516页)

Abstract: In the process of routing for data fusion in the wireless sensor network (WSN), it is necessary to minimize energy consumption. Due to the selfishness of node for conserving energy, it is likely to make every node refuse to transmit others data and energy consumption unevenly. In this case, the performance and lifetime of WSN is restricted. Based on the coalition game theory, we give an approach for modeling the above problem. We adopt the characteristic function to describe the gain and cost from a coalition. Then, we design a greedy algorithm to search the sub - optimal coalition structure in a large - scale solution space, and conclude that the solution is acceptable with a scope of errors. Experimental results show that the proposed can be well used to prolong the WSN lifetime.

Key words: wireless sensor network; data fusion; coalition game; characteristic function; sub - optimal coalitions structure