

一种近似 Markov Blanket 最优特征选择算法

崔自峰¹⁾ 徐宝文¹⁾ 张卫丰²⁾ 徐峻岭¹⁾

¹⁾(东南大学计算机科学与技术学院 南京 211189)

²⁾(南京邮电大学计算机学院 南京 210003)

摘 要 特征选择可以有效改善分类效率和精度,传统方法通常只评价单个特征,较少评价特征子集.在研究特征相关性基础上,进一步划分特征为强相关、弱相关、无关和冗余四种特征,建立起 Markov Blanket 理论和特征相关性之间的联系,结合 Chi-Square 检验统计方法,提出了一种基于前向选择的近似 Markov Blanket 特征选择算法,获得近似最优的特征子集.实验结果证明文中方法选取的特征子集与原始特征子集相比,以远小于原始特征数的特征子集获得了高于或接近于原始特征集的分类结果.同时,在高维特征空间的文本分类领域,与其他的特征选择方法 OCFS, DF, CHI, IG 等方法的分类结果进行了比较,在 20 Newsgroup 文本数据集上的分类实验结果表明文中提出的方法获得的特征子集在分类时优于其它方法.

关键词 特征选择;相关性;Markov Blanket;CHI-Square 检验;分类

中图法分类号 TP18

An Approximate Markov Blanket Feature Selection Algorithm

CUI Zi-Feng¹⁾ XU Bao-Wen¹⁾ ZHANG Wei-Feng²⁾ XU Jun-Ling¹⁾

¹⁾(School of Computer Science & Engineering, Southeast University, Nanjing 211189)

²⁾(School of Computer, Nanjing University of Posts & Telecommunication, Nanjing 210003)

Abstract Feature selection(FS) can effectively improve the speed and accuracy of classification. The traditional FS approaches usually score a single feature, do not evaluate feature subset. Based on the research in feature relevance, features can be further divided into four categories: Strong relevance, weak relevance, irrelevance and redundancy. The paper proposes a forward selection algorithm—An approximate Markov Blanket(MB) feature selection by theory of MB and Chi-Square test, which obtain an approximate optimal feature subset. Experiments on the datasets suggest that, compared with original feature set, the feature subset obtained by the proposed approach is much less than original feature set and performance on actual classification is better than or as good as that by original feature set. Meanwhile, when used in high dimension feature space such as text categorization, compared with other traditional feature selection approaches: OCFS, DF, CHI, IG, the performance obtained by the proposed method is obviously superior to that of others on 20 Newsgroup dataset.

Keywords feature selection; relevance; Markov Blanket; CHI-Square test; categorization

收稿日期:2006-05-11;最终修改稿收到日期:2007-03-05. 本课题得到国家杰出青年科学基金(60425206)、国家自然科学基金(60503020)和江苏省高校自然科学研究计划项目基金(04kjb520096)资助. 崔自峰,男,1976年生,博士研究生,主要从事信息检索、机器学习和模式分类等方面的研究. E-mail: maple.cui@126.com. 徐宝文,男,1961年生,博士,教授,博士生导师,主要从事程序设计语言、软件工程、并行与网络软件、知识与信息获取技术等方向的教学与科研工作. 张卫丰,男,1975年生,博士,副教授,主要研究方向为 Web 技术、信息检索、数据挖掘、Web 语义、人工智能、自然语言理解等. 徐峻岭,男,1984年生,博士研究生,主要从事信息检索、Web 挖掘等方面的研究.

1 引言

分类(又称有监督学习)是在给定一组带类标签的训练样例集合(通常以定长的特征向量表示)的情况下通过学习获得一个分类模型,反过来,再用学习到的模型预测未知事例的类标签.最近一段时间,相比于基于规则的学习算法,归纳学习算法得到了更多的研究者的青睐,因为归纳学习算法需要更少的时间消耗和人工参与.最常见的归纳学习算法有决策树学习^[1]、kNN 算法^[2]、贝叶斯学习^[2-3]和 SVM^[4]等.然而,人类知识的积聚造成许多归纳学习算法面临着“维数诅咒”(curse of dimensionality)的困难,即特征维数的增加使得归纳学习算法对时间和空间的需求也急剧上升,甚至复杂度呈指数级增长,同时,特征维数过高和训练样例不足常常导致归纳学习算法过度拟合训练样例的现象发生;而且,高维特征空间中存在着无关特征和冗余的特征更是导致问题的复杂化.

为了解决上述问题,特征选择是一种行之有效的办法.特征选择通常作为归纳学习前的一个预处理操作,它在问题的原始特征空间中选择一个最优的特征子集,使得对该子集上的操作可以很好地代表在原始特征空间上的操作.具体来说,特征选择可以去掉那些无关和冗余的特征,以此获得的特征子集用于分类可以减少所需的存储空间,加快处理速度以及提高分类精度. Dash 和 Liu^[5]等人对机器学习领域的特征选择作了深入研究.他们认为,在本质上,大多特征选择方法可以看作是一个搜索问题,每一个可能的特征子集作为搜索空间中的一个状态,根据对特征子集的操作,特征选择问题可以分成四个主要步骤:产生过程、评价函数、停止标准和校验过程.特征子集的产生过程又分为三种,即完全式、启发式和随机产生式.而按特征子集的评估函数划分为 5 种:距离度量、信息熵度量、依赖性度量、一致性度量和分类错误率度量^[5].

根据特征子集的产生过程是否依赖于最终使用它的归纳学习算法,特征选择方法可以分为两类: Wrapper 模型^[6]和 Filter 模型^[7]. Filter 模型选择特征的过程独立于具体的归纳学习算法,根据数据集内在本质特性来选择特征,没有继承归纳学习算法对特征的偏置; Wrapper 模型则使用具体归纳学习算法并以其性能作为评价和选择特征的标准,显然, Wrapper 模型方法继承了所使用的归纳学习算法偏置,只对预选的归纳学习算法有较好的性能.

Wrapper 模型方法的计算复杂度取决于学习算法,代价通常非常高昂,当特征数量很大时, Filter 模型是理想的选择.

现有的特征选择方法从特征评价方面又可以分成两种:单一特征选择评价和特征子集选择评价.单一特征选择评价方法是按照每一个特征在类中区分实例的重要性来评价特征^[8].特征子集选择评价方法是寻找满足某种重要性度量的最小特征子集,这种方法从整体角度来考虑,选择的特征子集通常比单一特征评价方法更优^[9-11].

目前,尽管特征选择的研究取得了许多重要的成果,但是仍然在文本分类和基因序列分析等领域面临困难,这些领域的共同特点是数据在高维空间分布.完全特征子空间搜索方法和 Wrapper 模型方法由于复杂性高无法应用于高维空间数据,而单一特征选择方法又难以获得合适的特征子集.本文的目标是研究高维特征空间的特征选择问题,并基于 Filter 模型提出一种新的特征子集评价的特征选择方法.通过形式化特征相关性分析,划分特征为相关、独立和冗余几个部分,进而通过 Markov Blanket(MB)理论与相关性和条件独立性度量获得一个近似最优的特征子集,相比于完全搜索算法只适用于特征较少的领域,本文方法计算复杂性最差情况下只有 $O(n^2)$, n 为特征维数,可以用于高维特征的选择.

文章下面首先讨论相关领域的工作;然后根据特征相关理论引出 MB 理论,证明通过 MB 得到相对冗余特征,进一步采用条件独立性度量给出近似 MB 定义并作为特征冗余的标准,在此基础上提出了一个近似 MB 的特征选择算法;最后给出详细的实验结果并与其它相关特征选择方法进行比较.

2 相关工作

在文本分类领域、信息论领域和统计理论中常用的特征选择方法有文档频率(DF)、信息增益(IG)、互信息(MI)、CHI-Square 检验(CHI)等^[8,11]. Yang 和 Pedersen 等人对目前常用于文本分类的 5 种特征选择方法做出了全面的比较研究,指出 DF、IG 和 CHI 获得了比 MI 和词条强度(TS)更好的结果^[8].然而,这 5 种方法在一定程度上只度量了特征和类之间的关系,忽略了特征之间依赖的关系,造成了特征冗余.这几种方法可以归结为基于 Filter 模型的单一特征选择.

单一特征选择虽然获得了较优的结果,但对于

高维特征空间仍存在不足. Yu 和 Liu 等人认为在高维空间中只消减那些无关的特征是不够的, 以此给出了一个新的特征选择框架, 在挑选出相关特征基础上, 用 IG 和信息熵方法来分析特征依赖性, 去掉冗余的特征^[9]. 基于同样的考虑, Qu 和 Hariri 等人提出一种基于信息论的 MI 的决策依赖和相关性分析^[10]. 特征子集的评价以决策相关性为奖励, 以两两特征的相关性为惩罚, 对特征子集打分, 评定是否高于给定阈值, 然而其实验部分的特征数只有 10 个, 没有对高维数据分析. 而且, MI 不完全服从度量性质, 可比性差, 文献[8]的实验结果表明, MI 方法在文本分类中的效果较差.

从最优子集的角度, Yan 和 Liu 等人提出在文本高维空间上基于正交质心点的最优特征选择^[12], 将特征选择问题映射到一个在离散解空间上的子空间的选择, 优化在子空间上的正交质心点的目标函数——类间散布矩阵, 来获得最优特征子集. 将线性代数领域中正交质心点算法应用于实数矩阵数据可以获得最优, 但是对文本词频矩阵进行计算在一定意义上反而不一定有效, 他们的实验结果证明该方法比传统 IG 和 CHI 方法改善并不是很大, 因为词频方法会造成较大的语义误差. 因此, 所谓的最优只是指在映射空间上的最优, 并不是原本问题的最优.

3 相关性分析、度量和算法

本文从特征相关性分析入手, 给出形式化定义, 建立起和 Markov Blanket 理论之间自然的联系, 得出相对冗余特征的概念, 通过 Markov Blanket 可以剔除特征集中的无关和冗余特征.

3.1 特征相关性分析

许多研究者对特征相关性做了较深入的研究, John, Kohavi 和 Pfleger 等人把整个特征空间划归为三类, 即强相关特征、弱相关特征和无关特征^[13]. 设 F 是特征集合, f_i 是一个特征, $S_i = F - \{f_i\}$, 三类特征的形式化的定义如下.

定义 1. 强相关. 特征 f_i 是强相关的当且仅当 $P(C | f_i, S_i) \neq P(C | S_i)$, 强相关特征是对类的分布构成影响的特征, 如果缺少了必然改变类的分布情况, 因此是最优子集的一部分.

定义 2. 弱相关. 特征 f_i 是弱相关的当且仅当 $P(C | f_i, S_i) = P(C | S_i)$ 且 $\exists S'_i \subset S_i$, 则 $P(C | f_i, S'_i) \neq P(C | S'_i)$,

弱相关特征在一定条件下影响类的分布, 但不一定是必须的.

推论 1. 无关性. 特征 f_i 是无关的当且仅当 $\forall S'_i \subseteq S_i, P(C | f_i, S'_i) = P(C | S'_i)$, 无关特征不影响类的分布情况, 是在分类中不需要的部分, 因此在特征选择中首先剔除.

特征子集评价一直是很困难的问题, 为了获得最优特征子集, Koller 和 Sahami 在统计理论上提出采用后向迭代删除过程获得最优特征子集, 删除过程中以特征是否存在 Markov Blanket 为标准, 故又称作 Markov Blanket 过滤. Markov Blanket 的定义以条件独立为基础, 由于文章篇幅所限, 本文直接给出其定义^[14].

定义 3. Markov Blanket. 给定一个特征 f_i , 设特征子集 $M_i \subset F (f_i \notin M_i)$, 称 M_i 是 f_i 的 Markov Blanket 当且仅当在给定 M_i 的条件下 f_i 和 $F - M_i - \{f_i\}$ 是独立的, 用公式表示如下:

$$P(F - M_i - \{f_i\} | f_i, M_i) = P(F - M_i - \{f_i\} | M_i).$$
 实际上 Markov Blanket 的定义不仅仅是关于其它特征的信息, 而且也包含类的信息. 从 Markov Blanket 和特征相关性的定义出发我们建立它们之间的联系, 得出如下结论.

推论 2. 如果特征子集 M_i 是 f_i 的 Markov Blanket, 那么在给定 M_i 的条件下 f_i 与类 C 也是独立的, 表示如下:

$$P(C | f_i, M_i) = P(C | M_i).$$
 从以上分析, 最优特征子集必然不包含无关特征, 弱相关会导致特征冗余, 如果一个特征 f 完全依赖于另一个特征, 那么可以看作相对于另一个特征冗余, 若不以另一特征为条件, 则它可能会是好的分类特征. 可见, 最优特征子集不可能包含相对冗余特征, 由此这里引入相对冗余的概念, 根据推论 2 给出相对冗余的定义.

定义 4. 相对冗余. 设 M 是一个特征集合, 如果特征 f_i 在 $M - \{f_i\}$ 中存在 Markov Blanket, 那么 f_i 相对于 M 来说是冗余.

定理 1. 设 G 是特征集合, 特征 $f_i (f_i \notin G)$ 在 G 中存在 Markov Blanket, 特征 $f_j (f_j \in G)$ 是相对于 G 的冗余特征, 从 G 中剔除特征 f_j 后, 特征 f_i 在 $G - \{f_j\}$ 中仍然存在 Markov Blanket.

定理 1 的意义在于当去除某些冗余特征后在开始阶段被剔除的特征仍然是冗余的. 该问题的证明涉及条件独立的相关性质^[15], 此处不再给出, 类似证明见参考文献[14].

从前面给出的定义、定理以及推论我们可以得

出下面结论.

定理 2. 强相关特征不存在 Markov Blanket.

证明. 用反证法. 假设强相关特征 f_i 存在 Markov Blanket M_i , 则根据推论 2 的公式与强相关定义相矛盾, 从而假设不成立, 得证. 证毕.

由此可见, 强相关特征不存在冗余, 是特征选择的最优特征子集中必不可少的一部分.

定理 3. 设 M_i 是 F 的特征子集, 特征 $f_i (f_i \notin M_i)$, 如果在给定 M_i 的条件下 f_i 与类 C 是独立的, 那么 M_i 是 f_i 的 Markov Blanket.

证明. 由已知条件: $P(C|f_i, M_i) = P(C|M_i) \Rightarrow P(f_i, M_i) = P(f_i|M_i)P(M_i) = P(M_i) \Rightarrow P(f_i|M_i) = 1$. 因此有 $P(F - M_i - \{f_i\} | f_i, M_i) = P(F - M_i - \{f_i\} | M_i)$, 根据 Markov Blanket 的定义可知 M_i 是 f_i 的 Markov Blanket. 证毕.

3.2 特征度量

为了度量特征之间的独立性, 有必要引入新的度量方式, CHI-Square 是检验变量之间独立性的一种方法. 最初 CHI-Square 检验的变量是服从贝努利分布, Pearson 进一步将其推广到多项分布^[16].

假设任意两个变量服从多项分布, 可以用一个 I 行 J 列的表来表示特征的分布, 表中每个表格 $Cell(i, j)$, $i = \{1, 2, \dots, I\}$, $j = \{1, 2, \dots, J\}$ 代表相应的事件, 对表格事件的发生情况进行计数, 计数值 n_{ij} 代表 $Cell(i, j)$ 事件发生的次数, 这样的表称作列联表. 设 $Cell(i, j)$ 表格事件发生的概率为

$$\pi_{ij}, i = \{1, 2, \dots, I\}, j = \{1, 2, \dots, J\},$$

列联表的第 i 行和 j 列的边缘概率为

$$\pi_{i\cdot} = \sum_{j=1}^J \pi_{ij}, \pi_{\cdot j} = \sum_{i=1}^I \pi_{ij}.$$

如果变量之间的分布是彼此独立的, 则有 $\pi_{ij} = \pi_{i\cdot} \times \pi_{\cdot j}$.

CHI-Square 检验是在不知道变量之间关系的情况下, 作如下的零假设检验:

$$H_0: \pi_{ij} = \pi_{i\cdot} \times \pi_{\cdot j}, i = 1, 2, \dots, I, j = 1, 2, \dots, J.$$

根据 CHI-Square 检验公式^[16]:

$$\chi^2 = \sum_{i=1}^I \sum_{j=1}^J \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \quad (1)$$

其中, O_{ij} 是样本观测值, E_{ij} 是 $Cell(i, j)$ 事件的期望值, χ^2 是一个服从 $(I-1)(J-1)$ 自由度的分布. 已知 O_{ij} 是 n_{ij} , 要求期望值 $E_{ij} = n \times \pi_{ij} = n \times \pi_{i\cdot} \times \pi_{\cdot j}$.

引理 1. 在 H_0 假设下, π_{ij} 的极大似然估计为

$$\hat{\pi}_{ij} = \hat{\pi}_{i\cdot} \times \hat{\pi}_{\cdot j} = \frac{n_{i\cdot}}{n} \times \frac{n_{\cdot j}}{n}.$$

引理 1 的证明由于篇幅问题不再给出, 相关知识

可见参考文献[16].

由引理 1, 表格 $Cell(i, j)$ 事件的期望 E_{ij} 用估计值代入: $\hat{E}_{ij} = n \times \hat{\pi}_{ij} = \frac{n_{i\cdot} n_{\cdot j}}{n}$.

式(1)可以改写为

$$\chi^2 = \sum_{i=1}^I \sum_{j=1}^J \frac{(n_{ij} - n_{i\cdot} n_{\cdot j} / n)^2}{n_{i\cdot} n_{\cdot j} / n} \quad (2)$$

一般来说, 统计假设检验通常采用临界值的表示方法, 这种方法需要根据专家知识事先确定显著性水平 α , 比如 $\alpha = 0.05$. 如果统计量 χ_0^2 (在不与 χ^2 混淆的情况下, 本文称 χ_0^2 为计算得到的统计量) 大于给定显著性水平的临界值 χ_α^2 , 那么就拒绝 H_0 . 这种方法的缺陷就是要求事先给定显著性水平 α , 为此本文采用 p -value 方法^[16], p -value 是可观测的显著性水平, 取值为大于统计量 χ_0^2 的概率分布值, 其中统计量 χ_0^2 服从 CHI-Square 分布.

定义 5. p -value (可观测的显著性水平) 是可以拒绝零假设 H_0 的最小显著性水平, 定义如下:

$$p_n(\chi_0^2) = \int_{\chi_0^2}^{\infty} \frac{1}{2^{n/2} \Gamma(n/2)} x^{(n/2-1)} e^{-x/2} dx, \quad x \geq 0, n \text{ 为自由度} \quad (3)$$

p -value 在本质上是概率值, 具有良好的度量性质 $0 \leq p\text{-value} \leq 1$. 如果 $p\text{-value} = 0$, 则表示以 100% 显著性水平拒绝零假设 H_0 , 也表示两变量之间依赖度高, 反之, $p\text{-value} = 1$ 表示变量之间是独立的. 使用 p -value 度量有几方面的优越性: 通过该值无需查分布表; 不必设定显著性水平 α , p -value 就是显著性水平; 用 p -value 做度量具有可比较性.

定义 6. 类相关性度量. 设 x 是特征, $C = \{c_1, c_2, \dots, c_i\}$ 是与 x 相关联的类别, 特征 x 与 C 的相关度 $R(x; C)$ 定义如下:

$$R(x; C) = 1 - P_1(\chi^2(x; C)),$$

$$\chi^2(x; C) = \sum_{i=1}^{|C|} pr(c_i) \chi^2(x; c_i) \quad (4)$$

定义 7. 条件独立性度量. 设 x 和 y 是两个特征, $x, y \in F$, 在不考虑类别信息时, 则称 $I_y(x, C)$ 为给定 y 的条件下 x 与 C 的独立度, 定义如下:

$$I_y(x, C) = P_1(\chi^2(x | y; C)),$$

$$\chi^2(x | y; C) = \sum_{i=1}^{|C|} \chi^2(x | y; c_i) \quad (5)$$

3.3 近似 Markov Blanket 算法

前面我们已经证明在分类过程中有两类特征是不需要的, 一类是无关特征, 另一类是弱相关且相对于其它特征子集是冗余的特征, 通过 Markov Blanket 方法我们可以去掉这两类不需要的特征, 得

到一个最优子集.但是,在高维特征空间中求 Markov Blanket 要对所有的特征子空间进行搜索,计算复杂性太高,其时间复杂性是 $O(2^n)$.可见对于像在文本分类领域,计算 Markov Blanket 几乎是不可能的.文献[14]对 Markov Blanket 集合基数大小进行了实验,认为在训练集合受限(不可能穷尽所有的文档)的情况下基数过大会导致过度拟合训练样本,当基数为 1 和 2 时就可以很好地逼近.因此,我们采用一种启发式的方法来近似 Markov Blanket.本文通过限定 Markov Blanket 的基数,根据定理 3 的结论来近似判定.

假定特征 f_i 和 f_j ,若有 $R(f_i; C) > R(f_j; C)$,则认为 f_i 包含的分类信息比 f_j 更强,判断 f_j 在给定的 f_i 条件下是否和类 C 独立,如果成立,那么认为 f_i 构成 f_j 的近似 Markov Blanket,进一步根据相对冗余的定义可得 f_j 相对于 f_i 冗余,可以从特征集合中剔除;否则, f_j 也是一个有效的分类特征.这里,我们采用启发性的条件独立性度量(式(5))作为判断 f_j 是 f_i 的近似 Markov Blanket 的标准,定义如下.

定义 8. 近似 Markov Blanket. 设特征 f_i 和 f_j ,且有 $R(f_i; C) > R(f_j; C)$, f_i 是 f_j 的一个近似 Markov Blanket 当且仅当 $I_{f_i}(f_j, C) > R(f_j, C)$.

由定理 1 可知,在前面阶段去掉的冗余特征在去除其它特征之后仍然是冗余的.然而对于近似 Markov Blanket 来说,该定理是不一定成立,但是对于强相关特征也不会存在近似 Markov Blanket,因此强相关特征是在任何阶段不会被剔除.总起来说,从构建一个强相关特征集 G 开始,若一个特征在 G 中不存在近似 Markov Blanket,则加入 G 中,对其余特征重复该过程,最后得到的子集合 G 就是一个近似最优特征子集.但是开始 G 如何选择呢?我们假设具有高的相关性的特征是强相关,因此把相关性最高的一个特征放入 G 中.

在以上分析的基础上,本文提出近似 Markov Blanket 特征选择算法,算法分两个部分,第一部分先选择相关特征集合,第二部分按照近似 Markov Blanket 的定义逐个去掉冗余的特征,最终获得一个近似最优特征子集.详细步骤见算法 1——近似 Markov Blanket 特征选择算法(AMB).在第 1 部分(1~6 行),AMB 首先根据相关性公式计算所有特征类的相关性,根据给定的相关性水平过滤所有无关的特征,得到一个相关特征集合 F ,然后按照相关性对 F 中的特征从大到小排序;第 2 部分(7~18

行),先初始化目标集合 G ,同时把第一个最相关特征加入 G ,然后从有序集合 F 中取下一个特征,在 G 中迭代寻找是否存在该特征的近似 Markov Blanket,如果不存在,则认为是一个强相关特征加入到目标集合 G ;否则,认为该特征为相对于 G 的冗余特征,继续取下一个特征,重复该过程直到 F 为空.

算法 1. 近似 Markov Blanket 特征选择算法(AMB).

输入: ρ //相关性水平

输出: G //特征子集

```

1. for each  $f_i$  of all features
    //过滤无关特征,选出相关特征子集  $F$ ;
2.   if  $R(f_i; C) > \rho$  then
3.     AddTo( $F, f_i$ );
4.   end if
5. end for
6. SortByDesc( $F$ ); //按相关性从大到小排序
7. Initialize  $G = \text{NULL}$ ; //目标集合  $G$ ,初始化为空
8.  $G \leftarrow f_i = \text{getFisrt}(F)$ ; //第一个最相关的特征到  $G$  中
9.  $f_j = \text{getNext}(F, f_i)$ ;
10. while ( $f_j < \text{NULL}$ )
11.   for each  $f_i$  in  $G$ 
12.     if  $I_{f_i}(f_j, C) > R(f_j, C)$  then
13.       skip  $f_j$  from  $F$ ; jump to 17;
14.     end if
15.   end for
16.    $G \leftarrow f_j$ ;
17.    $f_j = \text{getNext}(F, f_j)$ ;
18. end while

```

AMB 算法复杂性分析:算法需要计算特征的相关性和独立性,每一次计算的代价与样本个数相关,这里我们估计 AMB 算法关于特征个数上的复杂性,以一次相关性或独立性的计算为单位.假设特征个数为 n ,对所有 n 个特征的相关性都做计算,复杂性是线性 $O(n)$;判断一个特征是否存在近似 Markov Blanket,需要计算特征之间的独立性,如果目标集合中只选择了一个特征,其它皆是冗余的情况下,算法复杂性最好为 $O(n)$,在最差情况下(所有特征都不存在近似 Markov Blanket)算法的复杂性为 $O(n^2)$.一般情况下,AMB 算法的时间复杂度介于 $O(n)$ 和 $O(n^2)$ 之间.

4 实验分析

为了验证本文提出的方法,我们实现了归纳学习分类器和 AMB 算法以及几种不同的特征提取方法,

整个程序系统是用标准 C++ 实现,并在 Windows 2000 平台上运行通过.归纳学习分类器主要采用贝叶斯方法和决策树分类两种方法.整个实验的数据集分别采用机器学习领域中常用的数据集 Votes 和 Led24(噪声和无噪声)和文本分类领域的两个 20 Newsgroup 新闻组子集^①.实验结果表明本文提出的近似 Markov Blanket 方法选择的特征子集的分类性能优于其它方法.

表 1 机器学习常用数据集

数据集	类别数	原始特征数	AMB 提取	训练集合大小	测试集合大小
Votes	2	16	5	300	135
Led24_1(含 10%噪声)	10	24	13	2000(含 10%噪声)	100(无噪声)
Led24_2(无噪声)	10	24	16	2000(无噪声)	100(无噪声)

对于机器学习领域常用的数据集 Votes, Led24_1, Led24_2 分别采用 AMB 方法获得的特征子集数分别是 5, 13, 16 个特征,表 2 中详细列出特征子集,同时采用贝叶斯分类器和决策树分类器使用原始特征集合和 AMB 结果子集作分类测试,两种分类器的度量以分类正确率来评价,分类结果见表 2.对于 Votes 数据集合,AMB 算法结果子集仅选择了 30% 原始特征子集的特征数,最后获得的分类平均正确率为 95.285%,高于使用全部原始特征的分类正确率 93.565%. Led24 数据集是用数码管显示 10 个阿拉伯数字的特征,该数据集有 7 个特征是强相关

(1) 机器学习领域数据集
具体数据集的描述见表 1, Votes 数据集共有 435 个样本, 300 个用来训练,剩下的用作测试. Led24 是用程序根据用户的需要生成,生成过程中用户可以自己定义加入噪声比,我们一共生成了三个数据集: Led24_1(10% 噪声)和 Led24_2(无噪声),各含有 2000 个实例,用作训练集合,第 3 个用于测试的数据集含有 100 个实例(无噪声).

的,其它 17 个特征是随机生成的.分析选择出的特征子集,可以看出,由于引入噪声,完全相关的 7 个特征(0~6),只有 5 个(0~4)特征被选出来,影响了分类正确率.相比于无噪声数据集 Led24_2,完全相关的 7 个特征都被选择,分类正确率都达到了 100%.从表 2 我们可以看出决策树分类器对噪声敏感,因为 Led24_1 数据集特征选择前后决策树分类正确率只有 84%和 83%,而贝叶斯分类正确率都达到了 100%.由此可见,AMB 算法选择的特征子集以较少的特征,获得了优于或等于原始特征集合的分类正确率.

表 2 机器学习数据集实验结果

数据集	分类特征数	贝叶斯分类正确率/%	决策树分类正确率/%	平均正确率/%	AMB 算法结果子集
Votes	16/16	89.660	97.470	93.565	
Votes	5/16	94.480	96.090	95.285	0,2,3,14,15
Led24_1	24/24	100.000	84.000	92.000	
Led24_1	13/24	100.000	83.000	91.500	0~4,6,13,17~19,21~23
Led24_2	24/24	100.000	100.000	100.000	
Led24_2	16/24	100.000	100.000	100.000	0~7,11,13,16,18,19,21,22

(2) 文本分类高维特征数据集
本文方法主要是针对高维数据的特征选择,我们使用文本数据集 20 Newsgroup 作为我们的实验数据.两个文本数据集都来自 20 Newsgroup,为了减少文章中的噪声数据,对所有文档去掉头部的结构化标注说明,只包含文章的标题和内容,经过 Porter stemming 算法^[17](<http://www.tartarus.org/~martin/PorterStemmer>)作后缀统一化处理和去除 100 多常用词(stoplist).一个数据集(sport)来自娱乐和运动方面的 5 类: rec.sports.hockey, rec.sports.baseball, rec.autos, rec.motocycles, alt.altheism.经过 stoplist 列表过滤掉无意义词后

的原始特征数为 24084,然后进一步去掉文档频率小于 3 的特征,最后获得 11167 个特征的数据字典.另一个数据集(computer)也有 5 类都来自计算机领域: comp.graphics, comp.os.ms-windows.misc, comp.sys.ibm.pc.hardware, comp.sys.mac.hardware, comp.windows.x.去除无意义的词后的原始特征数为 44753 个,考虑的训练数据集的大小,删掉了文档频率小于 4 的特征,最终获得 7486 个特征的数据字典.

① Blake C, Merz C. UCI repository of machine learning databases, 1998. <http://www.ics.uci.edu/~mllearn/MLRepository.html>

表 3 20_Newsgroup 文本数据集

数据集	类别数	原始特征数	DF 过滤	AMB 提取	训练集合大小	测试集合大小
Newsgroup(computer)	5	44753	7486(4)	2595	4999	1504
Newsgroup(sport)	5	24084	11167(3)	2487	4000	1000

在文本分类领域里常用精度 P , 召回率 R 和 $Micro F1$ 对结果进行度量. 精度 P 是正确分类的数量比上测试数据的总数. 召回率 R 是正确分类的数量与预先标记的数据比率. 而 $Micro F1$ 度量结合了精度和召回率, 计算出一个综合的成绩, 计算公式如下:

$$Micro F1 = \frac{2(P \times R)}{P + R}.$$

Newsgroup(computer)数据集是文献[12]中使用的一组数据集, 为了比较, 我们选取了相同的数据集, 具有可比性. AMB 算法选择特征子集的总数为 2595 个, 从中分别取前面的 10, 100, 1000, 2595 个特征做实验, 与 OCFS, IG 和 CHI 方法进行比较, 见图 1 所示. 当特征数为 10 个时, AMB 算法的 $Micro F1$ 值仅小于 OCFS 方法, 优于 IG 和 CHI 方法. 当特征数大于 100 后, AMB 算法的 $Micro F1$ 值远高于其它三种方法, 由于 AMB 算法是子集的评价, 整个子集是近似最优的, 因此采用选择的子集 2595 个特征分类时, 达到最优的分类结果, 比 OCFS 的最高值 (10000 个特征) 高出 10%, 比 CHI 和 IG 方法高出 14%. 可见, AMB 算法以较少的特征获得了更好的分类结果.

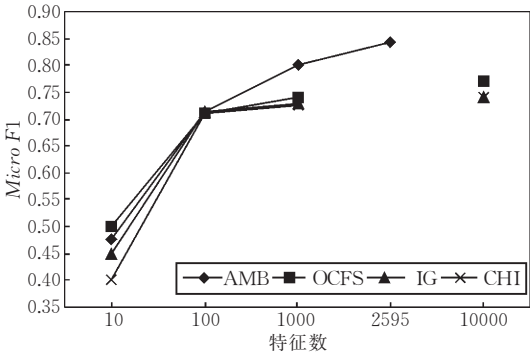


图 1 Computer 数据集上分类 $Micro F1$ 比较

在另一个文本数据集 Sport 上, 四种方法 AMB, IG, CHI 和 DF 作了比较, 如果如图 2 所示. 与 Computer 数据集的特点相似, 取 10 个 AMB 算法结果子集特征分类时, 其 $F1$ 值低于 CHI 方法的 $F1$ 值. 然而随着特征的增加, AMB 结果子集表现越好. 特征数为 1000 时, 四种方法难分仲伯. 不同的是, 整个 AMB 算法的结果子集 (2487 个特征) 用于分类时, 取得最好结果, 其它方法特征数在 10000 时却造成分类

效果不同程度的降低, 这也证明了特征维数过多, 含有大量无关特征和冗余特征, 不但导致分类算法的运算复杂度增加, 有时候还造成分类性能的下降.

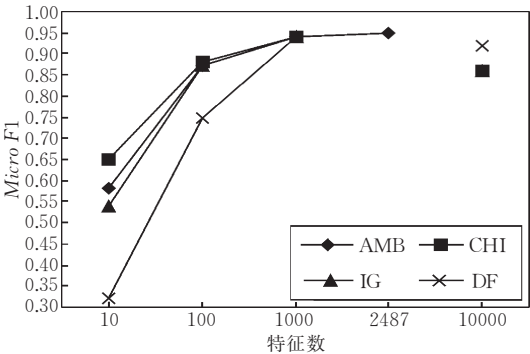


图 2 Sport 数据集上分类 $Micro F1$ 比较

综合以上两组数据集的实验结果可以看出, 传统的特征选择方法选择特征较少时, 确实能选择出部分较好的特征, 由于只注重单一特征的评价, 无法选择有最优的特征子集, 特征子集的数量也很难确定. 而 AMB 方法是在整个集合中寻找近似最优的特征子集, 获得的特征子集也最优地代表了整个特征集合空间, 以较少的特征获得了更优的性能.

在实验中, 我们还发现 IG 和 CHI 两种度量方法获得的特征排序非常类似, 因此他们的分类性能差距也非常小. DF 方法随着特征数量的增多, 分类性能有很好的表现, 在某些点上的分类性能和其它方法接近, 因此, 非常简单的方法也是不可忽视的.

5 总 结

本文提出了一种新的特征选择算法, 与传统的特征选择方法不同的是, 不仅仅评价单个特征, 更加注重特征子集的综合表现, 通过 Markov Blanket 的方法去掉冗余的特征, 得到近似最优的特征子集. 使用的独立性度量方法 Chi-Square 检验的优点是更加符合 Markov Blanket 的定义, 算法更精确. 实现过程中, 数据集特征的取值限制为二值分布特点, 文本数据特征也是二值假设, 即特征在文档中无论出现多少次, 计为 1, 不出现为 0. 从一定意义上说不利于文本分类, 但从分类结果看, 还是取得了不错的性能. 将来的工作可以在进一步扩大 Markov Blanket 子集的基数和度量方式两个方面深入研究.

参 考 文 献

[1] Mitchell T M. Machine Learning. New Jersey: McGraw Hill, 1997

[2] Duda R O, Hart P E, Stork D G. Pattern Classification. 2nd Edition. New York: John Wiley & Sons, 2000

[3] Rennie J D, Shih L, Teevan J, Karger D R. Tackling the poor assumptions of naive Bayes text classifiers//Proceedings of the 20th International Conference on Machine Learning. Washington DC, 2003: 616-623

[4] Joachims T. Text categorization with support vector machines: Learning with many relevant features//Proceedings of the 10th European Conference on Machine Learning. Chemnitz, DE, 1998: 137-142

[5] Dash M, Liu H. Feature selection for classification. International Journal of Intelligent data Analysis, 1997, 1: 131-156

[6] Kohavi R, John R C. Wrappers for feature subset selection. Artificial Intelligence, 1997, 97: 273-324

[7] Das S. Filters, wrappers and a boosting-based hybrid for feature selection//Proceedings of the 18th International Conference on Machine Learning. Williams College, 2001: 74-81

[8] Yang Y, Pedersen J O. A comparative study on feature selection in text categorization//Proceedings of the 14th International Conference on Machine Learning. Nashville, 1997: 412-420

[9] Yu L, Liu H. Efficient feature selection via analysis of rele-

vance and redundancy. Journal of Machine Learning Research, 2004, 10: 1205-1224

[10] Qu G, Hariri S, Yousif M. A new dependency and correlation analysis for features. IEEE Transactions on Knowledge and Data Engineering, 2005, 17: 1199-1207

[11] Wang G, Lochovsky F H, Yang Q. Feature selection with conditional mutual information maximization in text categorization//Proceedings of the 13th ACM Conference on Information and Knowledge Management. Washington DC, 2004: 342-349

[12] Yan J, Liu N, Zhang B, Yan S et al. OCFS: Optimal orthogonal centroid feature selection for text categorization//Proceedings of the ACM SIG on Information Retrieval. Salvador, Brazil, 2005: 122-129

[13] John G H, Kohavi R, Pfleger K. Irrelevant feature and the subset selection problem//Proceedings of the 11th International Conference on Machine Learning. New Jersey, 1994: 121-129

[14] Koller D, Sahami M. Toward optimal feature selection//Proceedings of the 20th International Conference on Machine Learning. Bari, Italy, 1996: 284-292

[15] Pearl J. Probabilistic Reasoning in Intelligent System. San Francisco: Morgan Kaufmann, 1988

[16] William M, Robert J B, Barbara M B. Introduction to Probability and Statistics (Reprint 11th ed.). Beijing: Machine Press, 2004

[17] Porter M F. An algorithm for suffix stripping. Program, 1998, 14(3): 130-137



CUI Zi-Feng, born in 1976, Ph.D. candidate. His research interests include information retrieval, machine learning and pattern classification etc.

XU Bao-Wen, born in 1961, Ph.D., professor, Ph.D. supervisor. His research interests include programming language, software engineering, parallel and network software,

acquisition technique on knowledge and information retrieval etc.

ZHANG Wei-Feng, born in 1975, Ph.D., associate professor. His main research interests include Web technologies, artificial intelligence, search engine, data mining and network language etc.

XU Jun-Ling, born in 1984, Ph.D. candidate. His research interests include information retrieval, and Web mining etc.

Background

The work in the paper is a part of the project "Feature Extraction in Email and Spam Email Recognition based on Agent", supported by the National Natural Science Foundation of China under grant of No. 60503020. Feature selection focus on finding as small feature subspace as possible to substitute for original space. Many induction learning methods do not have any loss on performance in subspace; meanwhile, feature selection can make those learning methods more effective and efficient, and save store space.

Now feature selection is facing challenge in high dimension space, such as text categorization in information retrieval. Many current popular methods are good at select single feature which has better capability to separate class than other features. However, most of the methods just consider the relation between feature and class, leaving behind the relation among features, which will result in redundancy. For this problem, the authors proposed a new approach based on Markov Blanket theory to remove the redundancy.