基于最小联合互信息亏损的最优特征选择算法

张逸石[1] 陈传波[1,2]

(华中科技大学软件学院 武汉 430074)[1] (华中科技大学计算机科学与技术学院 武汉 430074)[2]

摘 要 提出了一种基于最小联合互信息亏损的最优特征选择算法。该算法首先通过一种动态渐增策略搜索一个特征全集的无差异特征子集,并基于最小条件互信息原则在保证每一步中联合互信息量亏损都最小的情况下筛选其中的冗余特征,从而得到一个近似最优特征子集。针对现有基于条件互信息的条件独立性测试方法在高维特征域上所面临的效率瓶颈问题,给出了一种用于估计条件互信息的快速实现方法,并将其用于所提算法的实现。分类实验结果表明,所提算法优于经典的特征选择算法。此外,执行效率实验结果表明,所提条件互信息的快速实现方法在执行效率上有着显著的优势。

关键词 特征选择,条件互信息,最小联合互信息亏损,快速实现

中图法分类号 TP181 **文献标识码** A

Minimum Joint Mutual Information Loss-based Optimal Feature Selection Algorithm

ZHANG Yi-shi[1] CHEN Chuan-bo[1,2]

(School of Software Engineering, Huazhong University of Science and Technology, Wuhan 430074, China)[1]

(School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan 430074, China)[2]

Abstract In this paper, a minimum joint mutual information loss-based optimal feature selection algorithm was proposed, which firstly finds a non-discriminate feature subset of the original set via a dynamic incremental searching strategy, and then eliminates false positives by keeping minimum joint mutual information loss with class in each iteration using a minimal conditional mutual information criterion, in such a way as to obtain an approximate optimal feature subset. Furthermore, for the computationally intractable problem arising in high dimensional feature space that characterizes the existing method of conditional independence test with conditional mutual information, a fast implementation of conditional mutual information estimation was introduced and used to implement the proposed algorithm. Experimental results for the classification task show that the proposed algorithm performs better than the representative feature selection algorithms. Experimental results for the execution task show that the proposed implementation of conditional mutual information estimation has a considerable advantage.

Keywords Feature selection, Conditional mutual information, Minimum joint mutual information loss, Fast implementation

1 引言

特征选择已成为许多领域数据预处理特别是高维数据预处理过程中不可缺少的部分(例如文本分类、图像检索、生物信息处理等)。特征选择即从一组已知的特征集中按照某一准则选择出有很好的区分特性的特征子集,以达到用较少的特征对数据进行有效的表达和减少计算开销的目的。一般来说,特征选择方法可以分为3类[6]:嵌入方法、Filter方法和Wrapper方法。对于嵌入方法而言,特征选择被整合进某一特定学习算法的训练过程中,其典型代表是C4.5[11]。Filter方法基于某一具体的评价准则来选择特征,其选择特征的过程独立于具体的归纳学习算法,没有继承归纳学习算法对特征的偏置,其典型算法有ReliefF[12]等。与Filter方法相反,Wrapper方法以某一具体归纳学习算法的性能作为其评价和选择特征的标准。显然,Wrapper方法继承了其所使用的归纳学习算法对特征的偏置,因此只对预选的归纳学习算法有较好的性能,不具有一般性。此外,由于Wrapper方法使用了特定归纳学习算法作为评价标准,其计算代价通常非常高昂,因此当特征维数很高时,Filter方法往往要优于Wrapper方法。

常用的Filter型特征选择方法所采用的评价标准有卡方检验(χ^2-test)[10]、RMI[22]、信息熵[23]和互信息[8]等。然而,许多基于这些评价标准的特征选择方法仅考虑了单个特征与类标签之间的关系,未考虑特征之间的相关性,因此会造成特征冗余。为解决特征冗余问题,一系列基于条件独立性原理和马尔可夫毯(Markov blanket)理论的特征子集搜索算法被

到稿日期:2011-01-26 返修日期:2011-04-15 本文受国家自然科学基金项目(60973085)资助。

张逸石(1986-),男,硕士生,主要研究方向为模式识别,E-mail:easezh@126.com;陈传波(1957-),男,博士,教授,博士生导师,主要研究方向为模式识别与图像处理等。

相继提出[1,14,17]，这类算法通过对待选特征子集进行条件独立性测试与比较，进而搜索出一个与特征全集具有最小分类误差的近似最优特征子集。然而这类严格对特征子集进行条件独立性测试的算法普遍面临着“维数诅咒”的困难，其效率会随着特征子集规模的增长而急速降低[15]。解决该问题的工作可以归纳为以下3类：一类试图采用各自的搜索策略尽可能地减少搜索过程中条件独立性测试的次数[5,18,20]，然而它们并未降低条件独立性测试的计算复杂度，高维数据集下的效率瓶颈问题依然存在。另一类则对备选特征集的规模进行限制[1]，但这类方法往往导致结果精确性的降低。还有一类工作则采用某些近似条件独立性测试方法来替代对特征子集进行条件独立性测试（较为典型的有 Yu 等提出的 FCBF 算法[19]以及崔自峰等提出的 AMB 算法[21]等）。这种替代虽然有效地提高了运行效率，但对其替代策略的合理性无法给出理论上的证明，始终是其固有的缺陷。

为解决上述困难，本文以条件互信息作为评价准则，提出了一个基于最小联合互信息亏损的最优特征选择算法。算法首先通过一个动态渐增策略搜索最具分类能力且与当前已获得的特征子集独立性最强的特征，并确保在迭代结束时获得一个与特征全集具有相同联合互信息量的无差异特征子集。进而基于最小互信息原则在保证每一步都获得最小联合互信息亏损的情况下对可能的冗余特征进行筛选。本文还提出了一种对特征子集进行条件互信息估计的高效解决方法，并将其用于所提算法的实现。对 UCI 数据集、生物信息学领域的数据集以及 KDD Cup 中 Thrombin 数据集的分类实验表明，本文提出的特征选择算法性能要优于经典的 IG，ReliefF 和 I-AMB 特征选择算法。

2 联合互信息与条件互信息

联合互信息（joint mutual information）[17]用于描述两个变量集之间的相关性程度。令 $\boldsymbol{X}=\{X_1,X_2,\cdots,X_n\}$，$\boldsymbol{Y}=\{Y_1,Y_2,\cdots,Y_m\}$ 为两组有限离散型随机变量的集合，则 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 之间的联合互信息可通过分布 $p(\boldsymbol{x},\boldsymbol{y})$ 和 $p(\boldsymbol{x})p(\boldsymbol{y})$ 之间的 KL 散度（Kullback-Leibler divergence）[2]来表示：

$$\begin{aligned}I(\boldsymbol{X};\boldsymbol{Y})&=D_{KL}[p(\boldsymbol{x},\boldsymbol{y})\|p(\boldsymbol{x})p(\boldsymbol{y})]\\&=E_{p(x,y)}\left[\log\frac{p(\boldsymbol{x},\boldsymbol{y})}{p(\boldsymbol{x})p(\boldsymbol{y})}\right]\\&=\sum_{\boldsymbol{x}}\sum_{\boldsymbol{y}}p(\boldsymbol{x},\boldsymbol{y})\log\frac{p(\boldsymbol{x},\boldsymbol{y})}{p(\boldsymbol{x})p(\boldsymbol{y})}\end{aligned}$$

式中，log 表示以 2 为底的取对数运算（以下均使用 log 来表示 $\log_2$），$\boldsymbol{x}$、$\boldsymbol{y}$ 分别表示 $\boldsymbol{X}$、$\boldsymbol{Y}$ 中可能的取值组合。由联合互信息定义可知，若 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 中的变量有着紧密关联，$I(\boldsymbol{X};\boldsymbol{Y})$ 的值将会很大，反之其值将会很小。特别地，当 $I(\boldsymbol{X};\boldsymbol{Y})=0$ 时，$\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 相互独立。注意到当 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 分别仅含有一个变量时，即当 $\boldsymbol{X}=\{X\}$、$\boldsymbol{Y}=\{Y\}$ 时，$I(\boldsymbol{X};\boldsymbol{Y})=I(X;Y)$ 为变量 X 和 Y 的互信息。

令 $\boldsymbol{Z}$ 为一个有限离散型随机变量集合。将给定 $\boldsymbol{Z}$ 时 $\boldsymbol{X}$、$\boldsymbol{Y}$ 之间的条件互信息（conditional mutual information）定义为

$$\begin{aligned}I(\boldsymbol{X};\boldsymbol{Y}|\boldsymbol{Z})&=E_{p(z)}\{D_{KL}[p(\boldsymbol{x},\boldsymbol{y}|\boldsymbol{z})\|p(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{y}|\boldsymbol{z})]\}\\&=\sum_{\boldsymbol{z}}p(\boldsymbol{z})\Big(\sum_{\boldsymbol{x}}\sum_{\boldsymbol{y}}p(\boldsymbol{x},\boldsymbol{y}|\boldsymbol{z})\log\frac{p(\boldsymbol{x},\boldsymbol{y}|\boldsymbol{z})}{p(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{y}|\boldsymbol{z})}\Big)\\&=\sum_{\boldsymbol{z}}\sum_{\boldsymbol{y}}\sum_{\boldsymbol{x}}p(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z})\log\frac{p(\boldsymbol{x},\boldsymbol{y}|\boldsymbol{z})}{p(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{y}|\boldsymbol{z})}\end{aligned}$$

$I(\boldsymbol{X};\boldsymbol{Y}|\boldsymbol{Z})$ 用于描述在 $\boldsymbol{Z}$ 已知的情况下 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 之间的关联程度。若 $I(\boldsymbol{X};\boldsymbol{Y}|\boldsymbol{Z})=0$，则说明在给定 $\boldsymbol{Z}$ 时，$\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 条件独立；反之若该值越大，则说明在给定 $\boldsymbol{Z}$ 时 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 的关联程度越大。

联合互信息和条件互信息都具有非负性和对称性。其中对称性即 $I(\boldsymbol{X};\boldsymbol{Y})=I(\boldsymbol{Y};\boldsymbol{X})$，$I(\boldsymbol{X};\boldsymbol{Y}|\boldsymbol{Z})=I(\boldsymbol{Y};\boldsymbol{X}|\boldsymbol{Z})$。在信息论中，非负性和对称性是联合互信息和条件互信息最基本的性质。

3 基于联合互信息的最优特征选择

3.1 最优特征选择

这里使用机器学习的语境描述最优特征选择问题。给定一个训练集 $U=D(\boldsymbol{F},C)$，分类学习算法将会在 U 上训练出一个用于分类的映射规则 $h:\boldsymbol{F}\to C$。而最优特征选择的任务即从特征空间中选择出一组最小特征子集 $\boldsymbol{S}$，使得学习算法在该特征子集上建立的映射规则 $h':\boldsymbol{S}\to C$ 能够尽可能接近 $h:\boldsymbol{F}\to C$。由于难以直接比较映射规则的相似性，人们往往采用一定的评价函数来评价特征子集的优劣。一般而言，评价函数的值越大（或越小），其所评价的特征子集越优。令 $J(\cdot)$ 为特征子集的评价函数，最优特征选择的任务即在特征空间 $\boldsymbol{F}$ 中搜索出一个使特征选择算法所用评价函数 $J(\cdot)$ 的值尽可能大的最小特征子集 $\boldsymbol{S}$。

3.2 基于最大联合互信息的最优特选择方法

由于分类学习算法将会在给定训练集上训练出一个用于分类的映射规则，因此对样本集中每一个样本的分类，都存在一个从特征集合 $\boldsymbol{F}$ 到类标签 C 的映射与之对应。于是，当一个特征含有能够影响类标签分布的重要信息时，该特征即为相关特征，否则该特征为不相关特征或冗余特征[19]。鉴于联合互信息具有能够很好地对一组特征之间的信息关联进行描述的性质，它常被用作构造度量特征集合和类标签之间关联程度的评价准则。根据联合互信息的定义，$I(C;\boldsymbol{S})$ 的值越大，特征集 $\boldsymbol{S}$ 与类标签 C 的相关程度就越大，其所含有的分类信息就越多。反之，$I(C;\boldsymbol{S})=0$ 则表明 $\boldsymbol{S}$ 中不含与分类有关的信息。因此，若以联合互信息作为特征子集评价函数，即有 $J(\boldsymbol{S})=I(C;\boldsymbol{S})$，则最优特征选择任务可以表示为选取一个规模最小的特征子集 $\boldsymbol{S}$ 以满足 $\max_{\boldsymbol{S}\subset\boldsymbol{F}}I(C;\boldsymbol{S})$。设 $\boldsymbol{F}$ 为特征全集，$\{\boldsymbol{S},F\}\subset\boldsymbol{F}$，由联合互信息的链式规则[2]可知

$$I(C;F|\boldsymbol{S})=I(C;\boldsymbol{S},F)-I(C;\boldsymbol{S}) \tag{1}$$

由条件互信息的非负性知 $I(C;F|\boldsymbol{S})\geqslant 0$，于是有

$$I(C;\boldsymbol{S},F)\geqslant I(C;\boldsymbol{S}) \tag{2}$$

结合式(2)，可知 $\sup_{\boldsymbol{S}\subset\boldsymbol{F}}I(C;\boldsymbol{S})=I(C;\boldsymbol{F})$。为了讨论本文算法的搜索策略，以下首先给出无差异特征子集的定义。

定义 1（无差异特征子集） 设 $\boldsymbol{S}\subset\boldsymbol{F}$，若 $\boldsymbol{S}$ 满足

$$I(C;\boldsymbol{S})=I(C;\boldsymbol{F}) \tag{3}$$

则称 $\boldsymbol{S}$ 为 $\boldsymbol{F}$ 的一个无差异特征子集。

显然，最优特征子集是 $\boldsymbol{F}$ 的无差异特征子集之中规模最小的某一特征子集，这说明在 $\boldsymbol{F}$ 的无差异特征子集中有可能存在冗余特征。于是我们将特征选择问题分解成两个子问题，即首先得到一个 $\boldsymbol{F}$ 的无差异特征子集 $\boldsymbol{S}$，再对该子集进行缩减以排除冗余，从而完成特征选择任务。以下将分别针对

这两个子问题提出相应的搜索策略，并在此基础上给出本文的搜索算法。

3.3 基于最大条件互信息的无差异特征子集搜索策略

给定特征集 $\boldsymbol{S}$，若有 $I(C;F|\boldsymbol{S})>0$，则 F 是一个相关特征，且具有特征集 $\boldsymbol{S}$ 所不具有的关于类标签 C 的信息；若 F 与 C 条件独立，即 $I(C;F|\boldsymbol{S})=0$，则特征 F 在给定 $\boldsymbol{S}$ 时不具有关于分类的任何有用信息，即 F 是一个不相关特征，或者是一个在给定特征集 $\boldsymbol{S}$ 时的冗余特征。条件互信息 $I(C;F|\boldsymbol{S})$不仅考察了特征与类标签之间的相关性，同时考虑到了特征之间的相关性(冗余性判别)，因此它十分适合作为特征的评价准则。结合式(1)，我们采用一种动态渐增搜索策略，搜索一个 $\boldsymbol{F}$ 的无差异特征子集 $\boldsymbol{S}$。即在每一步迭代中，选择使当前条件互信息量最大的特征进入特征子集，并使用更新后的特征子集进行下一步的迭代。给定数据集 $U=D(\boldsymbol{F},C)$，设初始条件下待选特征集 $\boldsymbol{W}_1=\boldsymbol{F}$，已选特征集 $\boldsymbol{S}_1=\varnothing$，该迭代策略可以表示为

$$v(1)=\arg\max_{F\in\boldsymbol{W}_1} I(C;F|\varnothing)$$

$$\forall k,1\leqslant k<|\boldsymbol{F}|$$

$$v(k+1)=\arg\max_{F\in\boldsymbol{W}_{k+1}} I(C;F|F_{v(1)},\cdots,F_{v(k)})$$

式中，$F_{v(k)}$ 表示第 k 步时被选入特征子集的特征，$\boldsymbol{W}_{k+1}=\boldsymbol{W}_k-\{F_{v(k)}\}$。为避免每一次迭代后进行$I(C;\boldsymbol{S})\overset{?}{=}I(C;\boldsymbol{F})$的判断，以下给出一种等价的高效判别方法。

性质 1 令 $\sigma_k=\sum_{i=1}^{k-1}\max_{F\in\boldsymbol{W}_i} I(C;F|\boldsymbol{S}_i)$，$\boldsymbol{W}_i=\boldsymbol{F}-\boldsymbol{S}_i$，$\boldsymbol{S}_1=\varnothing$，则使 $I(C;\boldsymbol{S}_k)=I(C;\boldsymbol{F})$成立的充分必要条件是 $\sigma_k=I(C;\boldsymbol{F})$。

证明：根据联合互信息的链式规则，有

$$I(C;\boldsymbol{S}_i)=\max_{F\in\boldsymbol{W}_{i-1}} I(C;F|\boldsymbol{S}_{i-1})+I(C;\boldsymbol{S}_{i-1})(i=2,3,\cdots,k)$$

于是有

$$\begin{aligned}I(C;\boldsymbol{S}_k)&=\max_{F\in\boldsymbol{W}_{k-1}} I(C;F|\boldsymbol{S}_{k-1})+\max_{F\in\boldsymbol{W}_{k-2}} I(C;F|\boldsymbol{S}_{k-2})+\cdots\\&\quad+\max_{F\in\boldsymbol{W}_1} I(C;F|\boldsymbol{S}_1)\\&=\sum_{i=1}^{k-1}\max_{F\in\boldsymbol{W}_i} I(C;F|\boldsymbol{S}_i)=\sigma_k\end{aligned}$$

故有 $\sigma_k=I(C;\boldsymbol{F})\Leftrightarrow I(C;\boldsymbol{S}_k)=I(C;\boldsymbol{F})$。证毕。

性质 1 揭示了最大条件互信息累积量 σ_k 与 $I(C;\boldsymbol{S}_k)$之间的等价关系。由于$\max_{F\in\boldsymbol{W}_i} I(C;F|\boldsymbol{S}_i)$在每一步迭代中均已计算，因此使用 σ_k 进行 $I(C;\boldsymbol{S})\overset{?}{=}I(C;\boldsymbol{F})$的判断将大大提高算法的效率。此外，若在第 i 步中有$\max_{F\in\boldsymbol{W}_i} I(C;F|\boldsymbol{S}_i)=0$ 且 $\sigma_i\neq I(C;\boldsymbol{F})$的情况发生，为了保证策略的有效性，使搜索过程能够继续进行，我们选择当前最具有分类能力的特征(即和类标签具有最大互信息的特征)进入子集，即有 $\boldsymbol{S}_{i+1}=\boldsymbol{S}_i+\{F|\max_{F\in\boldsymbol{W}_i} I(C;F)\}$，$\boldsymbol{W}_{i+1}=\boldsymbol{W}_i-\{F|\max_{F\in\boldsymbol{W}_i} I(C;F)\}$，$\sigma_{i+1}=\sigma_i$。

3.4 基于最小条件互信息的冗余特征排除策略

上述搜索策略确保了在其迭代结束后所得特征子集是特征全集的一个无差异特征子集，同时确保了每次选入的特征最具有分类能力且不是当前特征子集的冗余特征，兼顾了冗余性的判别。其不足之处在于每一步迭代仅使用当前已获得的特征子集来评价新特征的分类能力及冗余性，对于早期进入特征子集的特征而言，它们的分类能力和冗余性并未被新进入的特征群所评价，于是存在早期进入特征子集的特征是随后进入特征子集的特征群条件下的冗余特征的可能性。由此便可能造成特征子集的规模过大，与特征选择算法搜索“最小特征子集”的目标相悖。因此，在获得 $I(C;\boldsymbol{S}_k)=I(C;\boldsymbol{F})$的特征子集 $\boldsymbol{S}_k$ 后，还需要用新特征群来评价老特征的方法筛选其中可能存在的冗余特征，从而减小特征子集的规模。另外，由于在实际数据集上估计条件互信息，往往会因训练样本有限而造成估计值与实际值之间存在偏差[13,15]。为减小该偏差所造成的干扰，可在本阶段中引入一个判别阈值 γ 来判断冗余性。即若有 $I(C;F|\boldsymbol{S}-\{F\})<\gamma$，则判定 F 为一个冗余特征并将其从特征子集中移除。此外，为进一步减小使用阈值进行冗余判别的错误率，同时使移除的特征对特征子集实际分类能力的影响尽可能小，还需使缩减后的特征子集与类标签之间联合互信息量的亏损尽可能小。为实现上述要求，采用一种基于最小条件互信息的迭代策略来筛选冗余特征。设算法在前一阶段所获得的特征子集为 $\boldsymbol{S}_k^0$，在本阶段中已经筛选出了 l 个冗余特征，则迭代流程可表示为

$$I(C;\boldsymbol{S}_k^{l+1})=I(C;\boldsymbol{S}_k^0)-\sum_{i=1}^{l}\min_{F\in\boldsymbol{S}_k^i} I(C;F|\boldsymbol{S}_k^i-\{F\}) \tag{4}$$

若在第 $l+1$ 步中有 $\min_{F\in\boldsymbol{S}_k^{l+1}} I(C;F|\boldsymbol{S}_k^{l+1}-\{F\})\geqslant\gamma$，则说明此时特征子集中的所有特征均不被判定为冗余特征。此时算法将退出本阶段的迭代，并输出最终所获得的特征子集 $\boldsymbol{S}_k^l$。

3.5 MJMIL 算法

综合上述讨论，我们给出 MJMIL 算法(Minimum Joint Mutual Information Loss-based algorithm)的伪代码。

算法 1 MJMIL 特征选择算法

Input: A *training dataset* $U=D(F,C)$
Output: Selected feature subset $\boldsymbol{S}$
1. Initialize: $S=\varnothing$, $\boldsymbol{W}=\boldsymbol{F}$, $\sigma=0$, JMI$=I(C;\boldsymbol{F})$
2. Repeat // 第一阶段
3. If $\exists F\in\boldsymbol{W}$ such that $I(C;F|\boldsymbol{S})>0$, do
4. Choose $F\in\boldsymbol{W}$ that maximizes $I(C;F|\boldsymbol{S})$
5. $\boldsymbol{S}=\boldsymbol{S}+\{F\}$, $\boldsymbol{W}=\boldsymbol{W}-\{F\}$, $\sigma=\sigma+I(C;F|\ \boldsymbol{S})$
6. Else
7. Choose $F\in\boldsymbol{W}$ that maximizes $I(C;F)$
8. $\boldsymbol{S}=\boldsymbol{S}+\{F\}$, $\boldsymbol{W}=\boldsymbol{W}-\{F\}$
9. End If
10. Until $\sigma=$JMI
11. Repeat // 第二阶段
12. Choose $F\in\boldsymbol{S}$ that minimizes $I(C;F\mid\boldsymbol{S}-\{F\})$
13. If $I(C;F\mid\boldsymbol{S}-\{F\})<\gamma$, do
14. $\boldsymbol{S}=\boldsymbol{S}-\{F\}$
15. End If
16. Until $\boldsymbol{S}$ *has not changed*

MJMIL 算法是一种基于特征子集条件独立性测试的算法，整个算法需对条件互信息进行 $O(|\boldsymbol{F}|^2)$次估计。然而在“条件”规模较大的情况下，条件互信息量的估计一直是最优特征选择算法和其他学习算法(例如贝叶斯网络构建算法等)的瓶颈问题之一[13,15]。当前已被明确提出并广泛应用于特征选择算法中的条件互信息估计方法，是一种时间复杂度与数据集样本规模呈线性关系同时与特征空间规模呈指数关系的方法[1,9]。受其限制，许多学习算法都不得不考虑限制“条件”的规模以保证算法的效率。以下将给出一种快速高效的条件互信息估计方法，该方法有效地解决了“条件”规模较大时效率低下甚至无法求解的困难，使 MJMIL 算法能够在高

维特征空间下高效地搜索最优特征子集。

3.6 条件互信息估计

在分析估计过程之前，首先给出局部互信息的定义以及条件互信息和局部互信息之间的关系。

定义 2(局部互信息) 给定数据集 $U=D(\boldsymbol{F},C)$，$X\in\boldsymbol{F}$，$Y\in\boldsymbol{F}$，$U'\subset U$。在 U 上估计的 X 和 Y 的局部互信息为

$$\hat{I}_{U'}(X;Y)=\sum_x\sum_y\hat{p}_{U'}(x,y)\log\frac{\hat{p}_{U'}(x,y)}{\hat{p}_{U'}(x)\hat{p}_{U'}(y)}$$

式中，$\hat{p}_{U'}(\cdot)$和$\hat{p}_{U'}(\cdot,\cdot)$分别为在 U' 上估计的概率和联合概率。

性质 2 对样本有限的离散型特征而言，其条件互信息的估计值可以表示为若干局部互信息的和。即给定一个含有 N 个样本的数据集 $U=D(\boldsymbol{F},C)$和特征子集 $\boldsymbol{S}\subset\boldsymbol{F}$，$X\in\boldsymbol{F}$，$Y\in\boldsymbol{F}$ 之间的条件互信息可表示为

$$\hat{I}(X;Y|\boldsymbol{S})=\sum_{i=1}^{q}\lambda_i\hat{I}_{U_i}(X;Y)\tag{5}$$

式中，$\lambda_i\in(0,1]$为一个对应于 U_i 的系数，$U_i\subset U$。

证明：设 $\boldsymbol{S}=\{F_1,\cdots,F_k\}$，$\Phi$ 为 $\boldsymbol{S}$ 中特征取值组合空间，$\phi_i\in\Phi$ 为 $\boldsymbol{S}$ 中特征的一种具体取值组合，Φ 中实际存在于 U 中的取值组合个数为 q，$U_i\subset U$ 为含有 ϕ_i 的子数据集，N_{U_i} 为 U_i 中的样本数，于是有 $\bigcup_{i=1}^{q}U_i=U$，且

$$\begin{aligned}\hat{I}(X;Y|\boldsymbol{S})&=\sum_{i=1}^{q}\hat{p}(\phi_i)\sum_x\sum_y\hat{p}(x,y|\phi_i)\log\frac{\hat{p}(x,y|\phi_i)}{\hat{p}(x|\phi_i)\hat{p}(y|\phi_i)}\\&=\sum_{i=1}^{q}\frac{N_{U_i}}{N}\sum_x\sum_y\frac{\pi_{U_i}(x,y)}{N_{U_i}}\log\frac{\frac{\pi_{U_i}(x,y)}{N_{U_i}}}{\frac{\pi_{U_i}(x)}{N_{U_i}}\frac{\pi_{U_i}(y)}{N_{U_i}}}\\&=\sum_{i=1}^{q}\frac{N_{U_i}}{N}\sum_x\sum_y\hat{p}_{U'}(x,y)\log\frac{\hat{p}_{U'}(x,y)}{\hat{p}_{U'}(x)\hat{p}_{U'}(y)}\\&=\sum_{i=1}^{q}\frac{N_{U_i}}{N}\hat{I}_{U_i}(X;Y)\end{aligned}$$

式中，$\pi_{U_i}(\cdot)$，$\pi_{U_i}(\cdot,\cdot)$分别表示某一特征取值和取值组合在 U_i 中出现的次数。令 $N_{U_i}/N=\lambda_i$，则有

$$\hat{I}(X;Y|\boldsymbol{S})=\sum_{i=1}^{q}\lambda_i\hat{I}_{U_i}(X;Y)$$

结合性质 2，可知估计条件互信息的运算时间等于 q 次局部互信息计算时间的总和。由于现有的互信息估计方法在一个含有 m 个样本的数据集上的时间复杂度为 $O(m)$[9]，因此若 q 个 U_i 在 U 中连续排列，则可以将其中每一个 U_i 视为一个独立的数据集，并且对 $\hat{I}_{U_i}(X;Y)$ 的计算即为在 U_i 上进行的估计互信息的计算，故此时估计条件互信息的时间复杂度为 $\sum_{i=1}^{q}O(N_i)$，即为 $O(N)$。为使 U_i 在 U 中连续排列，可采用以下两种解决方案：方案一，采用时间复杂度为 $O(N\log N)$ 的排序算法(例如归并排序等)，以每个特征 $F\in\boldsymbol{S}$ 为关键字对数据集 U 进行 $|\boldsymbol{S}|$ 次排序，然后估计所有的局部互信息并求和，最终得到条件互信息的值。此时整个条件互信息估计过程的时间复杂度为 $O(|\boldsymbol{S}|N\log N)+O(N)$；方案二，通过链式基数排序法对数据集 U 进行排序。设特征的取值上限为 r，由于链式基数排序法的时间复杂度为 $O(|\boldsymbol{S}|(N+r))$，于是最终求得条件互信息的时间复杂度为 $O(|\boldsymbol{S}|(N+r))+O(N)$，即为 $O(|\boldsymbol{S}|(N+r))$。可知当特征取值范围 r 不大时，特别对于二值数据集和稀疏数据集而言，方案二要优于方案一。本文采用方案二进行算法实现。

4 实验及结果分析

实验共分两个部分。第一部分，评价 MJMIL 算法性能。本次实验选择了经典的特征选择算法 IG[16]、ReliefF[12]以及 IAMB 算法[14]与 MJMIL 算法进行比较。分类器选择 Naïve Bayes[7]，kNN[8]和 C4.5[11]。第二部分，为验证本文提出的条件互信息求解算法的高效性，分别使用本文方法和文献[9]中的方法来实现 MJMIL 算法，并比较两种方法在 MJMIL 的前向迭代搜索阶段(第一阶段)上的执行时间。所有实验均在 2.8GHz CPU、2G RAM 的 PC 机上完成。

选取 UCI 库[1]中 4 个常用的基准数据集、生物信息学领域中的 Harvard Lung Cancer[2]数据集和 KDD Cup 2001 中的 Thrombin[3]数据集作为实验数据集。详细描述见表 1。

表 1 实验数据集描述

Datasets	Features	Instances	Classes
Kr-vs-kp	37	3196	2
Census-income	41	199523	2
Mfeat-factors	216	2000	10
Arrhythmia	279	452	16
Harvard Lung Cancer	12600	203	5
Thrombin	139351	2543	2

由表 1 可以看出，所选数据集无论是从样本数量(203～199523)，还是从特征域(37～139351)来说，都覆盖了相当广的范围，有利于全面检验算法的性能。

4.1 分类性能评价

采用著名的数据挖掘平台 Weka[4]作为分类实验平台。根据文献[12]，将 ReliefF 算法中的迭代参数设为 30，近邻个数设为 5。对于本文提出的 MJMIL 算法，将判别阈值 γ 设为 0.01。实验中 IAMB 算法[5]和 MJMIL 算法均由 Java 实现，并可嵌入 Weka 平台下使用。特征选择算法 IG 和 ReliefF 则可在 Weka 中直接调用。对于 Thrombin 数据集而言，由于其规模太大，我们用 C++实现了实验所用到的 4 种特征选择算法，并在 Weka 平台下对特征选择后的 Thrombin 数据集进行分类实验。实验采用分类准确率(Accuracy)作为分类性能的评价指标，并通过 10 次、10 折交叉验证法获得分类准确率。此外，为详细分析 MJMIL 算法在高维数据集上的性能，在使用分类准确率评价 Thrombin 数据集的同时，还使用 ROC 曲线面积(AUC)[3]、查准率(Precision)和查全率(Recall)作为评价指标。对于含有连续型特征的数据集，在实验

1) http://archive.ics.uci.edu/ml/

2) http://datam.i2r.a-star.edu.sg/datasets/krbd/LungCancer/LungCancer-Harvard1.html

3) http://pages.cs.wisc.edu/~dpage/kddcup2001/

4) http://www.cs.waikato.ac.nz/ml/weka/

5) 为使 IAMB 算法能够胜任于高维数据集，实验中该算法的条件独立性测试采用本文提出的实现方法进行实现

前先采用 MDL 方法[4]进行离散化处理。实验结果见表 2 和表 3。

表 2 所示实验结果表明，MJMIL 和 IAMB 算法在 3 种分类器上的平均分类准确率要明显优于 IG 和 ReliefF，且 MJMIL 的表现最优。对特征维数空间不高的数据集而言，除了在 Kr-vs-kp 上的表现外，MJMIL 和 IAMB 这两个涉及冗余特征处理的最优特征选择算法的优越性并不十分明显，IAMB更是在 Arrhythmia 数据集上获得了相对较差的效果，这在一定程度上说明了维数相对较低的数据集中特征冗余性对分类性能的影响并不是很大，此时特征与类标签的相关性对分类性能起到了决定性的作用。同时，IAMB 在 Arrhythmia 上所获得的较差结果也说明了 IAMB 算法在样本不足的数据集上容易错选较多的分类无关特征或冗余特征。对于具有中等以上规模特征维度的数据集直至 Thrombin 这样的高维数据集而言，MJMIL 则取得了较其他 3 种特征选择方法更加优越的结果，这说明了高维数据集中冗余特征的存在对分类结果有着重要的影响。筛选特征子集中冗余特征能够有效提高特征子集的质量，从而提高分类性能。在 MJMIL 和 IAMB 之间，MJMIL 的优越性更加明显，这表明了 MJMIL 算法具有更好的对相关特征搜索和对冗余特征判别的能力。

表 2 所选数据集上结合 4 种特征选择算法的分类性能比较(%)

	Kr-vs-kp	Census-income	Arrhythmia	Mfeat-factors	Harvard Lung Cancer	Thrombin	Avg. Accuracy
MJMIL+Naïve Bayes	**94.32**	**86.62**	**76.37**	**86.11**	**94.83**	**92.70**	**88.49**
MJMIL+kNN	**97.61**	94.02	68.87	**84.27**	**94.73**	**95.45**	**89.16**
MJMIL+C4.5	**97.85**	94.69	71.99	**80.31**	**91.67**	**95.48**	**88.67**
IAMB+Naïve Bayes	94.26	85.48	75.71	84.78	92.91	89.75	87.15
IAMB+kNN	97.13	94.08	68.94	83.02	93.25	91.27	87.95
IAMB+C4.5	97.26	**94.70**	71.50	78.85	90.25	90.19	87.13
IG+Naïve Bayes	88.09	79.99	74.05	82.68	92.27	92.41	84.92
IG+kNN	96.76	94.07	**69.36**	81.80	90.54	92.44	87.50
IG+C4.5	96.82	94.55	71.64	75.65	88.42	92.44	86.59
ReliefF+Naïve Bayes	91.34	81.60	73.81	59.90	92.76	91.25	81.78
ReliefF+kNN	96.75	**94.38**	67.30	64.86	90.59	94.34	84.70
ReliefF+C4.5	97.38	94.54	**72.88**	63.48	90.10	94.69	85.51

表 3 Thrombin 数据集上分类性能详细指标(%)

		MJMIL			IAMB			IG			ReliefF		
		Naïve Bayes	kNN	C4.5	Naïve Bayes	kNN	C4.5	Naïve Bayes	kNN	C4.5	Naïve Bayes	kNN	C4.5
ROC(AUC)		**92.62**	**85.92**	82.63	90.26	84.88	**83.58**	49.60	49.60	49.60	87.26	77.23	76.16
Precision	Active	**51.42**	**77.62**	**77.93**	39.76	76.19	76.98	0.00	0.00	0.00	44.19	71.30	71.09
	Inactive	97.17	96.63	**96.71**	**97.29**	**96.66**	96.46	92.41	92.44	92.44	96.59	95.47	95.82
Recall	Active	66.15	57.81	**58.85**	**68.75**	**58.33**	55.73	0.00	0.00	0.00	59.38	42.71	47.40
	Inactive	94.50	98.64	98.65	91.49	98.51	98.64	**99.96**	**100**	**100**	93.87	98.60	98.43

表 3 给出了算法在 Thrombin 数据集上的详细指标。注意到 3 种分类器在 IG 上的分类准确率分别是 92.41%，92.44%和 92.44%，然而其 AUC 值却都只有 49.6%；同时 3 种分类器与 IG 的组合在 Thrombin 上对 Inactive 类(2351 个样本)的查准率除了 Naïve Bayes 为 92.41%外，其余均为 92.44%(与分类准确率结果完全相同)；查全率除了在 Naïve Bayes 上为 99.96%外，其余均为 100%。但对于 Active 类(192 个样本)而言，其查准率和查全率全部为 0%，说明在进行交叉验证时，3 种分类器均将所有 Active 样本全部分类到 Inactive 类中。由此可见 IG 方法对 Thrombin 数据集的极端不平衡性(192:2351)十分敏感。而这个问题在 ReliefF 中则得到了一定的改善；在 IAMB 和 MJMIL 中，尤其是 MJMIL 中，得到了更好的改善。表 3 显示的 MJMIL 和 3 种分类器的 AUC 值分别为 92.62%，85.92%和 77.63%，除在 C4.5 上 MJMIL 的 AUC 值较 IAMB 略低以外，在 Naïve Bayes 和 kNN 上该值均要优于 IAMB。在分类准确率上，MJMIL 则全部优于另外 3 种特征选择算法。

4.2 条件互信息求解方法运行效率比较

为了更好地显示出本文提出的算法在高维数据集上的有效性，我们选择 Arrhythmia 和 Thrombin 数据集来测试算法的运行时间。为了进行比较，这里分别采用本文所提出的实现方法(Linear to the size of Samples and Linear to the size of Variables，记为 LSLV)和文献[9]提出的实现方法(Linear to the size of Samples but Exponential to the size of Variables，记为 LSEV)作为 MJMIL 算法中条件互信息量的求解方法。为增加度量的精确性，仅分析 MJMIL 算法的前向迭代搜索阶段(第一阶段)的执行时间。同时，由于 LSEV 方法对条件集合的规模有一定的限制(效率将随着条件集合的规模而急剧下降)，我们只给出 Arrhythmia 中前 10 个被选特征的执行时间，结果见图 1。

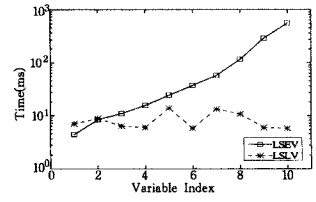

图 1 通过 LSLV 方法和 LSEV 方法实现的 MJMIL 在 Arrhythmia 数据集上第一阶段(部分)的执行时间比较

从图 1 可以看出，LSEV 方法的执行时间随着特征子集规模的增加呈指数增长，而 LSLV 方法的执行时间则几乎未随特征子集规模的增加而增长，仅呈现出一定范围内的波动。实验结果显示出 LSLV 方法较 LSEV 方法有着十分明显的优越性。为了更好地显示 LSLV 方法的效果，我们还给出了 LSLV 方法实现的 MJMIL 算法第一阶段在 Thrombin 数据

集上完整的执行时间。对于 LSEV 方法而言，由于其需要至少 2^{63} bit(MJMIL 第一阶段共选择 63 个特征，最少需 1 个 bit 存储 0 或 1)的存储空间，因此无法将其应用于 MJMIL 第一阶段在 Thrombin 数据集上的运行。实验结果见图 2。

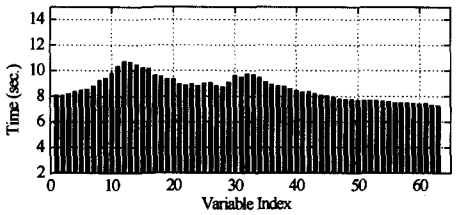

图 2 LSLV 方法实现的 MJMIL 在 Thrombin 上第一阶段的执行时间

从图 2 可以看到，MJMIL 算法在其第一阶段中一共选择了 63 个特征，对其中每一个特征而言，MJMIL 的处理时间均未超过 11s，且在第 44 个特征之后的所有特征执行时间均在 8s 及以下。此外，本次实验的总执行时间小于 600s，需要的内存空间也仅为 300M 左右(包括对 Thrombin 数据集的存储)，因此本方法对于一个含有 139351 个特征和 2543 个样本的巨型数据集而言，是一个相当有效的解决方案。实验表明，LSLV 方法成功地解决了基于特征子集的条件独立性测试所面临的高维瓶颈问题。

结束语 本文提出了一个最小联合互信息亏损的最优特征选择算法 MJMIL。该算法将最优特征选择问题分解成两个子问题：首先基于最大条件互信息评价方法通过一个动态渐增搜索策略获得一个特征全集的无差异特征子集；进而基于最小条件互信息原则在保证特征子集分类信息亏损尽可能小的同时，对可能的冗余特征进行筛选，以达到选择具有最优分类效果的特征子集的目的。在 Naïve Bayes、kNN 以及 C4.5上的分类实验结果证实了 MJMIL 算法的有效性及优越性。特别对高维数据集而言，MJMIL 算法能对高维特征域中的冗余特征进行有效筛选，分类实验结果明显优于经典的 IG、ReliefF 和 IAMB 特征选择方法。此外，本文还提出了一种“条件”中特征数量可变情况下条件互信息估计的快速实现方法。运行效率实验结果表明，此实现方法有效地解决了特征子集条件独立性测试面临的维数瓶颈，使本文提出的 MJMIL 算法能够胜任于高维数据集。

参 考 文 献

[1] Aliferis C, Statnikov A, Tsamardinos I, et al. Local causal and Markov blanket induction for causal discovery and feature selection for classification, part I: algorithms and empirical evaluation [J]. Journal of Machine Learning Research, 2010, 11: 171-234

[2] Cover T, Thomas J. Elements of Information Theory [M]. New York: Wiley, 1991

[3] Fawcett T. An introduction to ROC analysis [J]. Pattern Recognition Letters, 2006, 27(8): 861-874

[4] Fayyad U, Irani K. Multi-interval discretization of continuous valued attributes for classification learning [C]// Proceedings of the 13th International Joint Conference on Artificial Intelligence (IJCAI'93). San Francisco: Morgan Kaufmann, 1993: 1022-1027

[5] Fu S, Desmarais M. Fast Markov blanket discovery algorithm via local learning within single pass [J]. Lecture Notes in Artificial Intelligence, 2008, 5032: 96-107

[6] Guyon I, Elisseeff A. An introduction to variable and feature selection [J]. Journal of Machine Learning Research, 2003, 3: 1157-1182

[7] John G, Langley P. Estimating continuous distributions in Bayesian classifiers [C]// Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence (UAI'95). San Mateo: Morgan Kaufmann, 1995: 338-345

[8] Liu H, Sun J, Liu L, et al. Feature selection with dynamic mutual information [J]. Pattern Recgonition, 2009, 42(7): 1330-1339

[9] Margaritis D, Thrun S. Bayesian network induction via local neighborhoods [R]. TR-CMU-CS-99-134. 1999

[10] Qu G, Hariri S, Yousif M. A new dependency and correlation analysis for features [J]. IEEE Transactions on Knowledge and Data Engineering, 2005, 17(9): 1199-1207

[11] Quinlan R. C4.5: Programs for Machine Learning [M]. San Francisco: Morgan Kaufmann, 1993

[12] Robnik-Sikonja M, Kononenko I. Theoretical and empirical analysis of relief and reliefF [J]. Machine Learning, 53: 23-69

[13] Sotoca J, Pla F. Supervised feature selection by clustering using conditional mutual information based distances [J]. Pattern Recognition, 2010, 43(6): 2068-2081

[14] Tsamardinos I, Aliferis C, Statnikov A. Algorithms for large scale Markov blanket discovery [C]// Proceedings of the 16th International Florida Artificial Intelligence Research Society Conference (FLAIRS'03). Menlo Park: AAAI Press, 2003: 376-380

[15] Wang G, Lochovsky F. Feature selection with conditional mutual information maximin in text categorization [C]// Proceedings of the 19th ACM International Conference on Information and Knowledge Management (CIKM'04). New York: ACM Press, 2004: 342-349

[16] Yang Yi-ming, Pedersen J. A comparative study on feature selection in text categorization [C]// Proceedings of 14th International Conference on Machine Learning (ICML'97). Nashville: TN, 1997: 412-420

[17] Yang H, Moody J. Feature selection based on joint mutual information [C]// Proceedings of International ICSC Symposium on Advances in Intelligent Data Analysis. 1999: 22-25

[18] Yaramakala S, Margaritis D. Speculative Markov blanket discovery for optimal feature selection [C]// Proceedings of IEEE International Conference on Data Mining (ICDM'05). Washington, DC: IEEE Computer Society Press, 2005: 809-812

[19] Yu L, Liu H. Efficient feature selection via analysis of relevance and redundancy [J]. Journal of Machine Learning Research, 2004, 5: 1205-1224

[20] Zhang Yi-shi, Zhang Zi-gang, Liu Kai-jun, et al. An improved I-AMB algorithm for Markov blanket discovery [J]. Journal of Computers, 2010, 5(11): 1755-1761

[21] 崔自峰，徐宝文，张卫丰，等. 一种近似 Markov Blanket 最优特征选择算法[J]. 计算机学报，2007，30(12)：2074-2081

[22] 任永功，林楠. DPFS：一种基于动态规划的文本特征选择算法[J]. 计算机科学，2009，36(6)：188-191

[23] 宋国杰，唐世渭，杨冬青，等. 基于最大熵原理的空间特征选择方法[J]. 软件学报，2003，14(9)：1544-1550