基于互信息的输入变量选择算法综述

李香祯[1] 石晟玮[2] 殷小强[3] 刘 帅[4] 谢晓丹[1]

(1. 光学辐射重点实验室,北京,100854;2. 北京跟踪与通信技术研究所,北京, 100094;
3. 北京致生联发信息技术股份有限公司,北京, 100000;4. 66132 部队,北京, 100083)

摘要:空间环境是空间态势感知的重要组成部分,在空间科学领域,受限于对日地系统物理规律的全面认知,现有的基于物理机制的空间环境预测模型目前还难以实用化,通常采用的多是基于数据的经验类模型。基于互信息的输入变量选择算法为空间环境要素预测模型的输入确定了思路。由于能充分考虑不同输入、输入与输出之间的潜在关系,基于互信息的输入变量选择算法近年来在回归及分类问题上得到了广泛的应用和发展。本文以输入变量选择算法的 3 个关键环节,即评价标准、搜索策略和停止准则为线索,从不同角度对基于互信息的过滤式变量选择算法进行了系统的分析与梳理,重点对不同变量评价标准依赖的假设条件进行了数学上的推导和说明。最后总结了其发展规律,可为后续研究尤其是建立空间环境预测模型提供借鉴。

关键词:输入变量选择;互信息;回归模型;多步预测

中图分类号:TP391 **文献标志码**:A **文章编号**:1005-2615(2018)S2-0006-07

Overview of the Mutual Information-Based Input Variable Selection Method

LI Xiangzhen[1], *SHI Shengwei*[2], *YIN Xiaoqiang*[3], *LIU Shuai*[4], *XIE Xiaodan*[1]

(1. Science and Technology on Optical Radiation Laboratory, Beijing, 100854, China;
2. Beijing Institute of Tracking and Telecommunitions Techonology, Beijing, 100094, China;
3. Zhisheng Lianfa Information Technology Co. Ltd, Bejing, 100000, China;
4. 66132 Army, Beijing, 100083, China)

Abstract: The space environment is an important part of space situational awareness. In the field of space science, due to the incomplete understanding of the physics laws of the solar-terrestrial system, some existing space environmental prediction models based on physical mechanisms are still difficult to practicalize. It is more common to use the data-based empirical models. The input variable selection(IVS) algorithm based on mutual information(MI) can help to determine the impact inputs of the spatial environmental prediction models. Since taking into account of the potential relationship between different inputs, and between inputs and outputs, the MI-IVS algorithms have been widely applied to the regression and classification problems in recent years. From different perspectives, this paper systematically analyzes the current widely-used filter-based MI-IVS algorithms with the three key points of the IVS algorithms, namely the evaluation criteria, the search strategy and the stopping criterion as a clue. It focuses on the mathematical derivation of the assumptions of these criterias. Finally, the trend of the MI-IVS algorithms is summarized, which can provide reference for subsequent research, especially for establishing space environmental prediction models.

Key words: input variable selection; mutual information; regression model; multi-step prediction

空间环境是空间态势感知的重要组成部分。 而空间环境与大多数的自然环境系统一样,具有典

收稿日期:2018-03-23;**修订日期**:2018-05-30

通信作者:李香祯,女,高级工程师,E-mail:18911650912@163.com。

引用格式:李香祯,石晟玮,殷小强,等. 基于互信息的输入变量选择算法综述[J]. 南京航空航天大学学报,2018,50(S2):6-12. LI Xiangzhen, SHI Shengwei, YIN Xiaoqiang, et al. Overview of the mutual information-based input variable selection method[J]. Journal of Nanjing University of Aeronautics & Astronautics,2018,50(S2):6-12.

型的非线性、非平稳特征，在空间科学领域，受限于对日地系统物理规律的全面认知，现有的基于物理机制的空间环境预测模型目前还难以实用化，因此，在日常业务中，通常采用的多是基于数据的经验类模型。

为了尽可能准确的描述物理现象，在建立预测模型时，需要构造较大的备选输入变量集合，尤其在多元时间序列分析中，较大的延迟项意味着指数级增长的变量集合。当输入变量的样本量为无穷大或相对于变量规模足够大时，越多的输入变量意味着越强的描述能力。但实际应用中，受限于分析方法和计算条件，输入变量多并不直接意味着预测结果更好，甚至使分析结果变差。因而，需要对输入变量进行必要的选择，增加有限样本集时相关估计量的置信度。

互信息能够描述变量之间的线性与非线性关系，且对数据的统计分布没有任何约束，使得基于互信息的变量选择方法得到了非常广泛的研究与应用。本文从数据出发，借鉴信息论中的互信息概念，对基于互信息的输入变量选择算法进行系统梳理和分析。

1　基于互信息的输入变量选择算法概述

输入变量选择的关键步骤有[1-2]：确定评价标准、制定搜索策略和停止准则。评价标准用于衡量待选变量子集的优劣，搜索策略用于生成待选变量子集，停止准则用于确定最优或次优变量子集的维度。其流程图见图 1。

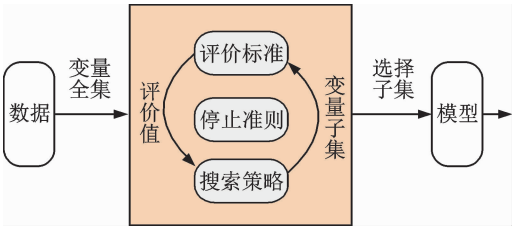

图 1　过滤式变量选择算法流程图

Fig. 1　Filtering variable selection algorithm flow

利用互信息、条件互信息等概念及其变化形式，可以得到各种各样的变量选择标准。文献[3，4]对此进行了专题的综述，试图将这些方法统一到共同的框架内。本文以其为基础，从变量选择的整个流程出发，涉及标准的估计、参数的确定、搜索策略的分析等，探讨不同标准建立的基本原理和假设条件，进一步丰富了变量选择标准的理论基础，确保作为方法或算法的完整性。按照评价标准单次评价的是单个变量还是变量子集，可以将变量选择标准划分为单变量选择标准和分组选择标准；按照搜索路径的范围可以将搜索策略划分为局部搜索策略和全局搜索策略[2]；停止准则分为事先确定变量子集的维度和自动确定变量子集维度两种。后文按此思路分别予以介绍。

2　变量选择标准

在变量选择问题中，通常需要对变量间的关系进行定义，包括输入变量与输出之间的相关关系，输入变量之间的冗余关系和辅助关系。在互信息背景下，对变量间关系可以描述为[5-6]：

强相关关系：对于任意变量 x_i，称其与输出 $\boldsymbol{Y}$ 存在强相关关系，如果条件互信息 $I(x_i;\boldsymbol{Y}\mid \boldsymbol{X}_{\setminus x_i})>0$。

弱相关关系：对于任意变量 x_i，称其与输出 $\boldsymbol{Y}$ 存在弱相关关系，如果条件互信息(1) $I(x_i;\boldsymbol{Y}\mid \boldsymbol{X}_{\setminus x_i})=0$；且(2) $\exists \tilde{\boldsymbol{X}}\subset \boldsymbol{X}$，使得 $I(x_i;\boldsymbol{Y}\mid \tilde{\boldsymbol{X}})>0$。

统计独立关系：对于变量 x_i，称其与输出统计独立，如果对于 $\forall \tilde{\boldsymbol{X}}\subset \boldsymbol{X}_{\setminus x_i}$，$I(x_i;\boldsymbol{Y}\mid \tilde{\boldsymbol{X}})=0$。

冗余关系：对于变量 x_i，称其为冗余变量，如果其满足(1)x_i 是弱相关变量；(2)，使得 $I(x_i;\{\boldsymbol{Y},\tilde{\boldsymbol{X}},\boldsymbol{X}_{\setminus x_i}\}\mid \tilde{\boldsymbol{X}})=0$。

辅助关系：对于变量 x_i，称其与变量集合 $\boldsymbol{X}$ 关于输出 $\boldsymbol{Y}$ 存在辅助关系，如果

$$I(x_i;\tilde{\boldsymbol{X}};\boldsymbol{Y})=I(x_i;\boldsymbol{Y}\mid \tilde{\boldsymbol{X}})-I(x_i;\boldsymbol{Y})>0$$

文献[3]以已选变量子集 $\boldsymbol{S}$ 作为条件项，定义变量的相关性，即 $I(x_i;\boldsymbol{Y}\mid \boldsymbol{S})>0$，但忽略了待选变量与已选变量子集存在弱相关关系的情况。文献[4]指出强相关的条件过于严格，但其所举证明实例实际属于弱相关关系。

强相关表明待选变量包含了其他变量不包含的输出信息；弱相关是一种广义上的冗余关系，待选变量可以被其他变量(集)替换而不影响对输出的描述；冗余则表示待选变量所包含的所有信息，包括输出、补集等，都能被另一个变量(集)所表示，能够被完全的替换；辅助关系表示多个变量(集)的联合能够提供比各自单独作用时更多的输出信息量；统计独立关系则描述的是待选变量不包含任何关于输出的信息，也不存在任何辅助变量。

那么对于基于互信息的过滤式变量选择方法而言，比较合理的定义为[4,7-8]：变量子集称为最优变量子集，当变量子集所包含的关于输出变量的信息与变量全集包含的关于输出变量的信息等量，且维度最少。本质上就是要尽可能的保留相关变量，去除冗余变量和无关变量。由于实际中很难做到对高维互信息的准确估计，Guyon 等[7]指出可以通过优化问题近似得到最优解集。

2.1 单变量选择标准

单变量选择标准就是每次只能评价单个变量与输出及已选变量子集之间的关系,选择最优的变量加入到当前最优变量子集中。

2.1.1 高维互信息选择标准

直接利用高维互信息进行变量选择,能够将多个变量之间的交互作用考虑在内,更容易获得全局最优的变量子集。

Bonev 等[9]利用熵图法对高维互信息进行估计,进而提出了直接利用高维互信息进行变量选择的评价标准。

Chow 等[10]首先利用聚类算法实现对样本数据的压缩,再利用剪枝 Parzen 窗估计算子结合二次互信息进行进行高维互信息估计,提出了直接基于高维互信息估计的变量选择算法。

2.1.2 低阶近似选择标准

在面对有限样本集和高维变量时,在一定假设条件下,用低维互信息近似高维互信息是一种计算效率和计算精度的折中,可以有效避免所谓的维数灾难问题。

2.1.3 Mm(Maximum of minimum)类标准

不失一般性情况下,对任意均值始终位于最小值与最大值之间。从近似的角度,最小值是对均值的保守估计,最大值则是对均值的乐观估计。对于变量选择,保守意味着捕获的细节更多,从而筛选掉更多的变量;而乐观意味着选择更多的变量,但可能是无关或冗余变量。根据变量选择的目标,一般会采用最小值或直接用均值来表示变量子集的互信息水平。

上述标准中都包含有变量之间以及变量与输出之间的求和或平均项,一定程度上平滑了变量间的关系,使得结果包含有部分冗余变量。因此,考虑用待选变量与已选变量及输出之间相互关系的最小值而不是均值来表示待选变量的相关程度,此即 Mm 类标准的基本思想。

2.1.4 归一化类标准

在对含有较多取值的变量进行互信息估计时,会使得互信息向其产生偏差。因此,一些学者尝试对变量选择标准加入归一化的项,如 Normalized mutual information feature selection (NMIFS)[11]是对 mRMR 的改进,Double input symeter relevance (DISR)[12]是对 JMI 的改进。

文献[13]考虑已选变量集合中任意两个变量的组合与待选变量之间的关系,对 NMIFS 提出了改进,即 NMIFS-2 标准。

综合以上分析结果可知,局部变量选择标准都是在一定的假设条件下,利用低阶互信息对高阶互信息近似所得的结果。不同的假设条件会得到不同的的不同选择标准,体现了对变量间相关、冗余因素的不同考虑。

局部变量选择标准的优点是计算效率高,是一种计算效率和精度的良好折中。缺点是由于只考虑了一阶、二阶近似,不能充分体现高维变量子集之间及其与输出之间的交互关系,是一种局部最优解。

2.2 分组变量选择标准

分组变量选择标准的出发点为:直接评估某个变量子集的优劣,待选变量以组合的形式出现。此时,从初始变量全集中直接选取变量子集。仍然采用低维互信息近似高维互信息,得到分组变量选择标准。

与局部变量选择标准不同,全局变量选择标准直接衡量整个变量子集的优劣,其优点是:只要生成足够的变量子集,全局式变量选择标准更容易获得全局最优解或近似最优解。其缺点亦很直观:由于涉及变量子集内所有的变量参与运算,计算量较大。

3 搜索策略

搜索策略与变量选择标准有着深刻的联系。一般来说,分为逐个选取变量的局部搜索策略、选取整个变量子集的全局搜索策略,以及近年来出现的以聚类分析为代表的其他类策略。

3.1 局部搜索

局部搜索策略就是逐个评估待选取的输入变量,选取满足评价标准的当前最优变量,是一种贪心搜索策略。按照搜索方向的不同,一般分为前向、后向和综合式搜索策略。前向搜索是一种增量式的搜索过程,包括顺序前向选择法(Sequential forward selection, SFS)、顺序前向浮点式选择法、步进式前向搜索方法等,主要与单变量选择标准组合使用;后向搜索是逐步的去除不相关及冗余变量,是一种减量式搜索,包括顺序后向删减法(Sequential backward elimination, SBE)、后向浮点式删减法等,主要与分组变量评价标准组合使用;综合式是将两者进行结合,包括前向-后向特征选择方法、增 r 减 l 式特征选择方法等[1-2,14]。本文重点介绍比较经典的 SFS、SBE,他们也是其他搜索策略的基础。

(1)顺序前向选择法(SFS)。该类方法初始时刻的变量子集为空集。由于所有前向搜索方法的过程都是增量式的,都需要处理第一个变量子集的选取问题。通常的做法是从输入变量全集中选取单个或几个变量,计算其与输出变量之间的互信

息，选取最大值对应的输入变量构成初始变量子集，即

$$\boldsymbol{S}_1=\{x^*:x^*=\max_{x_i\in X}(I(x_i;\boldsymbol{Y}))\} \tag{1}$$

选定第一个变量子集后，按照前面介绍的变量选择标准依次选取接下来的变量

$$\boldsymbol{S}^{t+1}=\boldsymbol{S}^t\bigcup\operatorname*{argmax}_{x_i\in\boldsymbol{X}-\boldsymbol{S}^t}J(x_i) \tag{2}$$

(2)顺序后向删减法(SBE)。该类方法的初始变量子集为包含所有变量的全集。确定初始变量子集后，结合变量选择标准逐步从已选变量中去除不相关或冗余变量

$$\boldsymbol{S}^{t+1}=\boldsymbol{S}^t-\operatorname*{argmax}_{x_i\in\boldsymbol{S}^t}J(x_i) \tag{3}$$

贪心式搜索能够取得较高的计算效率，但其问题也相对突出，主要表现在所谓的嵌套效应，即一旦某个标量被选中(增加或删除)，其都不能在后续的搜索过程中被改变，虽然综合式搜索能够有效的缓解这一情况，但仍容易陷入局部最优。

3.2　全局搜索策略

当确定全部的待选变量之后，变量选择问题本质上是一个多目标优化或带有约束项的优化问题。因而，全局寻优算法逐渐受到重视，比较典型的有遍历搜索法(如分支界定法[15])、随机式搜索算法(如遗传算法[16-17]、蚁群算法[18]、粒子群优化算法[19-21]等)和凸优化算法如二次规划算法[22-24]等。全局搜索策略一般与分组变量选择标准组合使用。

(1)遍历搜索法，即穷尽搜索法。该方法将变量组合的所有情况进行遍历，能够真正意义上获得训练数据集上的最优输入变量子集。设待选输入变量为 P 个，则共有 2^P-1 个非空变量子集，当 P 取值较大时，此方法变的不可取。

(2)随机式搜索算法。此类算法将全部变量构成的集合视为搜索空间，按照某种规则在此空间中随机的游走寻找最优解。由于该类方法多是仿生类算法，易于理解和编程实现，有较高的寻优效率和精度，因此应用范围越来越广泛。除了不同算法的不同动力学特性外，其共同的步骤有两个：①确定编码规则；②确定适应度函数。

按照变量选取与否的二值特性，一般直接将变量的组合优化问题编码为{0,1}字串，1 代表该变量被选中，0 表示未被选中[20,25]。由此形成遗传算法中的染色体编码和粒子群优化中的粒子位置编码等。此时，问题的解将在离散空间中进行寻找。有时为了利用某些算法的连续性特征，也可以将粒子编码为连续值，再将其映射为选取变量的字串。以粒子群优化算法为例，设定一个阈值 δ，当更新后的粒子位置 $x_i>\delta$ 时，第 i 个变量在字串中的符号编码为 1，表示被选中，否则编码为 0 表示未被选中[20,26-27]。

随机搜索算法在变量选择中的应用以封装式选择方法居多，适应度函数(或评价函数)通常将回归模型的预测精度与模型的复杂度相结合，得到一个综合指标。在过滤式选择算法中，受限于互信息的估计问题，随机式搜索算法应用较少，比较典型的主要有：Xue Bing[21]在其博士学位论文中系统的研究了粒子群优化算法与互信息在变量选择中的应用和改进，并将变量选择问题转化为多目标优化问题；与之类似，Karakaya 等[28]利用 Borg multiobjective evolutionary algorithm (MOEA)作为搜索算法，以变量子集的相关度、冗余度、子集规模及分类器精度为目标函数，提出了能够获得具有最优精度的一系列解集的变量选择算法。

随机式搜索算法有较高的寻优效率，能够搜索较大的解空间，但由于随机性，得到的可能是近似最优解[7]。

(3)凸优化类算法。将变量选择形式化为凸优化问题，利用优化算法进行求解早有研究[29]。近几年，基于二次规划的变量选择算法再次成为研究热点。在二次规划问题的目标函数中，用二次项描述输入变量之间的相似关系，用线性项描述输入变量与输出之间的相关关系，由于二次规划问题在组合优化领域已经得到了很好的发展，具有较为完整的理论基础，因此在变量选择领域的研究和应用也会受到越来越多的关注。

3.3　聚类搜索策略

从聚类的角度，可以将变量间的相关和冗余关系对应为类内相似度最大、类间相似度最小，从每个类别中找到代表性的变量表征此类别，再对聚类后的代表性变量进行如上的变量选择工作。此即聚类变量选择算法的基本思想。基本步骤包括两步：(1)按照一定标准对初始变量集合进行聚类；(2)从各类别中选出代表性变量；(3)确定最终的变量子集。第 3 步可以按照前述变量子集选取策略执行。

文献[30]以条件互信息为基准，提出了变量间距离的度量如下

$$\boldsymbol{D}(x_i;\widetilde{x}_i)=\frac{I(x_i;\boldsymbol{Y}\mid\widetilde{x}_i)+I(\widetilde{x}_i;\boldsymbol{Y}\mid x_i)}{2H(\boldsymbol{Y})} \tag{4}$$

式中，$\widetilde{x}_i$ 表示类别的代表性变量，x_i 表示任意变量。

各类别的代表性变量选取为与输出最相关的变量，即 $x_i^*=\{x_k:\max_{x_k\in X_i}I(x_k;\boldsymbol{Y})\}$，$x_i^*$ 表示第 i 个类别的代表变量。

文献[30]证明了 $D(x_i;\widetilde{x}_i)$ 满足非负性、三角

不等式、对称性等距离度量的基本属性，但文献[31]指出其证明所依赖的假设条件存在问题。文献[32]将其应用到针对连续型变量的单输出和多输出回归预测问题中，从应用角度证实了方法的有效性。

文献[33]以归一化的互信息(Symmetric uncertainty, SU)为基准，其度量指标为

$$\mathrm{SU}(X;Y)=\frac{2\times I(X;Y)}{H(X)+H(Y)} \tag{5}$$

各类别的代表性变量同样选取为与输出最相关的变量，此时 $x_i^*=\{x_k:\max_{x_k\in \boldsymbol{X}_i}\mathrm{SU}(x_k;\boldsymbol{Y})\}$。

4 停止准则

变量选择的目标是尽可能的去掉无关和冗余变量，使得选出的输入变量子集在保证精度的同时，具有最少的变量规模。

多维互信息单调非减，意味着不能仅仅通过判断输入变量所带来的互信息增减来确定最终输入变量子集的维度，并且，互信息估计算子的偏差也能导致互信息一定程度的非理想变化。比如，增量搜索时，随着变量子集的规模增加，互信息会逐渐增大，但有些变量可能是冗余变量；在全局搜索策略中，变量子集中可能含有冗余、无关变量，但全局搜索策略并不能将此因素考虑在内。换句话说，前面的变量选择标准和搜索策略，都没有对变量子集的规模进行限制，不同规模的变量子集可能由于上述因素会产生相同的变量选择标准值，使得结果出现偏差。

因此，需要对变量子集的规模进行限制。通常的做法有两种：(1)事先指定要选择的变量数目，当达到此数目时即停止[2,8,30]，或者在维度为定值时，选取变量选择标准最佳值对应的各个变量；(2)根据一定的准则自动确定变量规模，如前后变量选择标准的变化程度满足一定条件即停止[34-36]，或满足一定标准的最小规模变量集合[16-20]。

5 结束语

本文从数据驱动的角度出发，研究了在建立空间环境预测模型时确定输入变量集合的方法，即基于互信息的输入变量选择算法，通过对多种算法的分析和对比，阐述了部分算法的适用条件，从理论上研究了该类方法在空间环境建模中的适用性，为后续建模工作提供了参考。基于互信息的输入变量选择算法虽然在近年得到了很好的发展，但仍然存在严峻的挑战：

(1)关于互信息在变量选择中应用的合理性需要进一步的理论探讨。互信息以其能够反映变量之间的非线性关系、对数据分布特征的零约束为特点被采纳到变量选择标准中，但其绝对大小并不能直接对应变量间相关性的强弱，文献[37]对互信息在分类问题中变量选择的适用性进行了讨论，指出并不是所有情况互信息都能够适用。因此，需要从理论和实验上深入探讨互信息描述变量间关系的合理性和适用条件。

(2)关于互信息的估计算法需要进一步的研究。虽然基于样本间距离的多维互信息估计，从准确性和稳定性方面都要优于核函数法和柱状图法等，但当变量空间规模非常大或(和)样本规模非常大时，如何有效的确定估计算子的参数将变得困难；估计算子不仅要处理离散型变量，还要处理连续型变量，甚至两种变量同时存在，此时对估计算子形成了新的挑战，文献[38,39]对此类问题进行了初步的研究。

(3)建立基于互信息的变量选择算法统一框架，包括理论框架和实验框架。从文中即可看出，基于互信息的变量选择标准有多种变体，且多是基于研究者的巧妙设计得到的结果，并在不同的数据集(部分是在公共数据源，但最优的变量子集数目未知)针对不同问题进行的研究，其通用性不能得到保证。由于基于互信息的变量选择标准出发点是一致的，文献[3]在建立统一的理论框架方面做了很好的尝试，但仅针对的是局部变量选择标准。文献[30]建立了适用于环境科学的统一实验框架，其数据符合环境科学的基本特性，且最优输入变量子集已知，便于对新算法的比较测试，但缺少更加通用、更具代表性的数据和测试工具。

(4)从优化的角度深入研究变量选择算法。将变量选择问题形式化为优化问题后，利用凸优化或多目标优化求解算法[21, 31-43]进行求解将在变量选择领域得到越来越多的重视。多目标优化算法可提供用户多个可供选择的解集，能够得到更加符合实际应用的输入变量；凸优化求解算法具有充分的理论基础，能够获得较为一致的全局最优或近似最优解，当前主要以二次规划问题为主，但关系矩阵的正定性等条件仍有待进一步完善。此外，适合于高维数据处理的聚类变量选择算法近年来也逐渐得到重视。

(5)引入新的反映变量间关系的度量标准。关于互信息的研究，在信息论中有深刻的理论背景，但近年在统计领域出现了一些新的能够描述非线性关系、适用于多维向量的度量标准，如距离相关系数[44]、Heller-Heller-Gorfine (HHG)测度[45]等，其不需要额外的参数，且计算过程仅涉及样本间距离的计算，可进一步分析与实验新标准在变量

选择领域的适用性。

参考文献：

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

[2] MAY R, DANDY G, MAIER H. Review of input variable selection methods for artificial neural networks[G]// Artificial neural networks-methodological advances and biomedical applications. Croatia: InTech, 2011: 19-44.

[3] BROWN G, POCOCK A, ZHAO M, et al. Conditional likelihood maximisation: A unifying framework for information theoretic feature selection[J]. Journal of Machine Learning Research, 2012, 13(1): 27-66.

[4] VERGARA J, ESTÉVEZ P. A review of feature selection methods based on mutual information[J]. Neural Computing & Applications, 2014, 24(1): 175-186.

[5] MEYER P, SCHRETTER C, BONTEMPI G. Information-theoretic feature selection in microarray data using variable complementarity[J]. IEEE Journal of Selected Topics in Signal Processing, 2008, 2(3): 261-274.

[6] KOJADINOVI I. Relevance measures for subset variable selection in regression problems based on k-additive mutual information[J]. Computational Statistics & Data Analysis, 2005, 49(4): 1205-1227.

[7] GUYON I, ELISSEEFF A. An introduction to feature extraction[J]. Studies in Fuzziness and Soft Computing, 2006, 207: 1-25.

[8] BATTITI R. Using mutual information for selecting features in supervised neural net learning[J]. IEEE Transactions on Neural Networks, 1994, 5(4): 537-550.

[9] BONEV B, ESCOLANO F, CAZORLA M. Feature selection, mutual information, and the classification of high-dimensional patterns[J]. Pattern Analysis & Applications, 2008, 11(3/4): 309-319.

[10] CHOW T, HUANG D. Estimating optimal feature subsets using efficient estimation of high-dimensional mutual information[J]. IEEE Transactions on Neural Networks, 2005, 16(1): 213-224.

[11] ESTÉVEZ P A, TESMER M, PEREZ C A, et al. Normalized mutual information feature selection[J]. IEEE Transactions on Neural Networks, 2009, 20(2): 189-201.

[12] MEYER P, BONTEMPI G. On the use of variable complementarity for feature selection in cancer classification[J]. Lecture Notes in Computer Science, 2006, 3907: 91-102.

[13] SHUANG C, YU H. Mutual information based input feature selection for classification problems[J]. Decision Support Systems, 2012, 54(1): 691-698.

[14] NOVOVICOVA J, PUDIL P, SOMOL P. Efficient feature subset selection and subset size optimization[C]// Pattern Recognition Recent Advances. Croatia: InTech, 2010: 4-1-4-23.

[15] NARENDRA P, FUKUNAGA K. A branch and bound algorithm for feature subset selection[J]. Electronics Letters, 1989, c-26(9): 917-922.

[16] GE H, HU T. Genetic algorithm for feature selection with mutual information[C]// Seventh International Symposium on Computational Intelligence and Design. [S. l.]: IEEE, 2014, 1: 116-119.

[17] HUANG J, CAI Y, XU X. A hybrid genetic algorithm for feature selection wrapper based on mutual information[J]. Pattern Recognition Letters, 2007, 28(13): 1825-1844.

[18] ZHANG C, HU H. Ant colony optimization combining with mutual information for feature selection in support vector machines[J]. Lecture Notes in Computer Science, 2005, 3809: 918-921.

[19] 姚旭，王晓丹，张玉玺，等. 基于粒子群优化算法的最大相关最小冗余混合式特征选择方法[J]. 控制与决策，2013，28(03)：413-417.
YAO Xu, WANG Xiaodan, ZHANG Yuxi. A maximum relevance minimum redundancy hybrid feature selection algorithm based on particle swarm optimization[J]. Control & Decision, 2013, 28(3): 413-417.

[20] XUE B, ZHANG M, BROWNE W. Particle swarm optimisation for feature selection in classification: Novel initialisation and updating mechanisms[J]. Applied Soft Computing, 2014, 18(4): 261-276.

[21] XUE B. Particle swarm optimisation for feature selection in classification[D]. Wellington: Victoria University of Wellington, 2014.

[22] VINH N, CHAN J, ROMANO S, et al. Effective global approaches for mutual information based feature selection[C]// Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. [S. l.]: ACM, 2014: 512-521.

[23] RODRIGUEZ-LUJAN I, HUERTA R, ELKAN C, et al. Quadratic programming feature selection[J]. Journal of Machine Learning Research, 2010, 11(2): 1491-1516.

[24] BOUAGUEL W, MUFTI G B. An improvement direction for filter selection techniques using information theory measures and quadratic optimization[J]. International Journal of Advanced Research in Artificial Intelligence, 2012, 1(5): 7-10.

[25] ZHANG Y, WANG S, PHILLIPS P, et al. Binary PSO with mutation operator for feature selection using decision tree applied to spam detection[J]. Knowledge-Based Systems, 2014, 64: 22-31.

[26] CHUANG L, KE C, YANG C. Boolean binary particle swarm optimization for feature selection[C]// IEEE World Congress on Computational Intelligence. [S. l.]:IEEE, 2008:2093-2098.

[27] ZHANG Y, GONG D, HU Y, et al. Feature selection algorithm based on bare bones particle swarm optimization[J]. Neurocomputing, 2015, 148:150-157.

[28] KARAKAYA G, GALELLI S, AHIPASAOGLU S, et al. Identifying (Qquasi) equally informative subsets in feature selection problems for classification: A max-relevance min-redundancy approach[J]. IEEE Transactions on Cybernetics, 2015, 46(6): 1424-1437.

[29] BRADLEY P, MANGASARIAN O, STREET W. Feature Selection via Mathematical Programming[J]. Informs Journal on Computing, 1998, 10:209-217.

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

[31] VINH N , BAILEY J. Comments on supervised feature selection by clustering using conditional mutual information-based distances[J]. Pattern Recognition, 2013, 46(46):1220-1225.

[32] CARMONA P, SOTOCA J, PLA F. Filter-type variable selection based on information measures for regression tasks[J]. Entropy, 2012, 14(2): 323-343.

[33] SONG Q, NI J, WANG G. A fast clustering-based feature subset selection algorithm for high-dimensional data[J]. IEEE Transactions on Knowledge and Data Engineering, 2013, 25(1): 1-14.

[34] VINH N, CHAN J, BAILEY J. Reconsidering mutual information based feature selection: A statistical significance view[C]//Twenty-Eighth AAAI Conference on Artificial Intelligence. [S. l.]:[s. n.], 2014.

[35] SHARMA A. Seasonal to interannual rainfall probabilistic forecasts for improved water supply management: Part 1-A strategy for system predictor identification[J]. Journal of Hydrology, 2000, 239(1): 232-239.

[36] FRANÇOIS D, WERTZ V, VERLEYSEN M. The permutation test for feature selection by mutual information[C]// 14th European Symposium on Artificial Neural Networks. Belgium:[s. n.], 2006:239-244.

[37] FRÉNAY B, DOQUIRE G, Verleysen M. Theoretical and empirical study on the potential inadequacy of mutual information for feature selection in classification[J]. Neurocomputing, 2013, 112(10):64-78.

[38] 王皓, 孙宏斌, 张伯明. PG-HMI:一种基于互信息的特征选择方法[J]. 模式识别与人工智能, 2007, 20(01):55-63.
WANG Hao, SUN Hongbin, ZHANG Boming. PG-HMI: Mutual information based feature selection method[J]. Pattern Recognition & Artificial Intelligence, 2007, 20(01):55-63.

[39] ROSS B. Mutual information between discrete and continuous data sets.[J]. Plos One, 2014, 9(2): e87357-1-5.

[40] GALELLI S, HUMPHREY G, MAIER H, et al. An evaluation framework for input variable selection algorithms for environmental data-driven models[J]. Environmental Modelling & Software, 2014, 62:33-51.

[41] WANG Z, LI M, LI J. A multi-objective evolutionary algorithm for feature selection based on mutual information with a new redundancy measure[J]. Information Sciences, 2015, 307:73-88.

[42] KAMYAB S, EFTEKHARI M. Feature selection using multimodal optimization techniques[J]. Neurocomputing, 2015, 171:586-597.

[43] HOQUE N, BHATTACHARYYA D, KALITA J. MIFS-ND: A mutual information-based feature selection method[J]. Expert Systems with Applications, 2014, 41(14):6371-6385.

[44] SZEKELY G, RIZZO M, BAKIROV N. Measuring and testing dependence by correlation of distances[J]. Annals of Statistics, 2007, 35(6): 2769-2794.

[45] HELLER R, HELLER Y, GORFINE M. A consistent multivariate test of association based on ranks of distances[J]. Biometrika, 2012, 100(2): 503-510.

(编辑:张蓓)