

# 特征降维技术的研究与进展

黄 铨

(西南交通大学信息科学与技术学院 成都 610031)

**摘要** 数据特征的质量会直接影响模型的准确度。在模式识别领域,特征降维技术一直受到研究者的关注。随着大数据时代的到来,数据量巨增,数据维度不断升高。在处理高维数据时,传统的数据挖掘方法的性能降低甚至失效。实践表明,在数据分析前先对其特征进行降维是避免“维数灾难”的有效手段。降维技术在各领域被广泛应用,文中详细介绍了特征提取和特征选择两类不同的降维方法,并对其特点进行了比较。通过子集搜索策略和评价准则两个关键过程对特征选择中最具代表性的算法进行了总结和分析。最后从实际应用出发,探讨了特征降维技术值得关注的研究方向。

**关键词** 降维,特征选择,特征提取,研究进展

中图分类号 TP391 文献标识码 A

## Research and Development of Feature Dimensionality Reduction

HUANG Xuan

(School of Information Science and Technology, Southwest Jiaotong University, Chengdu 610031, China)

**Abstract** Quality of data characteristics directly impacts the accuracy of the model. In the field of pattern recognition, dimensionality reduction technique is always the focus of researchers. At the era of big data, massive data needs to be processed while the dimension of the data is rising. The performance of the traditional methods of data mining is degraded or losing efficiency for processing high dimensional data. Studies show that dimensionality reduction technology can be implemented to effectively avoid the “Curse of Dimensionality” in data analysis, thus it has wild application. This paper gave detailed description about two dimensionality reduction methods which are feature selection and feature extraction, in addition, a thoroughly comparison about the feature of these two methods was performed. Feature selection algorithm was summarized and analyzed by two key steps of algorithm, which are searching strategy and evaluation criterion. Finally, the direction for future research of the dimensionality reduction was discussed based on its practical application.

**Keywords** Dimensionality reduction, Feature selection, Feature extraction, Research progress

## 1 引言

在机器学习和模式识别领域,数据特征一直备受研究者关注。数据特征质量的好坏会直接影响模型的效果。随着数据采集方式和存储技术的不断发展,人们已具备获取和存储更多信息的能力;特别是近年来物联网、云计算等技术的迅速发展,使得数据信息快速地渗透到了生活中的各个领域,大数据时代来临<sup>[1]</sup>。这些庞大的数据为研究者们提供了更丰富、更完整的信息,但同时越来越巨大的数据量和逐渐增加的维度也给研究者的后续工作带来了困难。高维数据中包含大量的冗余特征,这些冗余特征会对后续的数据分析过程产生干扰,从而影响最后的结果;同时还会影响算法的复杂度,增加计算开销。若用传统的学习方法来处理高维数据,会使算法的学习性能降低甚至失效,从而导致数据资源成了“维数灾难”。如何利用这些大数据并从中挖掘出对后续算法最有效的特征,以解释数据中蕴含的有效信息,使其更容易被人们理解,同时避免陷入“维数灾难”,是目前的研究热点。

对于“维数灾难”,目前有两种常用的处理方法:1)针对特

定的高维数据重新设计新的算法来处理数据;2)先对原始数据进行降维处理,再利用降维后的数据特征来进行后续分类或聚类学习。本文主要讨论第二种方法。图 1 给出原始数据的特征降维过程,即首先将原始数据进行预处理,以去除信息中的干扰和噪声;然后通过线性或者非线性的转换,或从原始特征中直接选取少量有效的特征,在不损失原有信息的前提下将数据从高维约减到低维;最后,将降维后的数据用于分类或者聚类。特征降维技术不仅可以提高后续算法的分类精度和泛化性能,还能降低算法的复杂度,使后续数据分析更加容易。

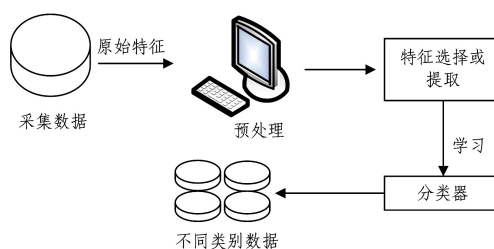


图 1 数据模式识别系统

特征降维主要分为特征选择和特征提取两个方面。

## 2 特征提取

特征提取是通过线性或非线性转换将原始数据的特征映射到低维空间,同时在低维空间中保持大量原始数据相关信息的过程<sup>[2]</sup>。特征提取技术在计算机视觉、图像处理等领域已被广泛应用。近几十年来,研究者们已经提出了许多特征提取方法,根据是否线性,将这些方法分为线性特征提取和非线性特征提取两个类。

### 2.1 线性特征提取

在很长一段时间内,机器学习中的模型主要是线性的,它们都是在假设数据服从全局线性的条件下建模且模型中的变量相互独立。典型的线性特征提取方法主要有:主成分分析(PCA)<sup>[3]</sup>、线性判别分析(LDA)<sup>[4]</sup>、独立成分分析(ICA)<sup>[5]</sup>等。

PCA 是人们最熟悉的降维方法,它的基本思想是将数据特征映射到新的特征空间中,并将其转化为几个综合特征,这些综合特征是原特征的线性组合,它们最大化地保留了原特征的信息,且彼此间不相关。PCA 擅长处理呈高斯分布的原始数据,通过协方差的度量最大限度地去除噪声和冗余,但它不能用于空间非线性的数据。现实世界中,数据分布较复杂,远超出了高斯分布,因此 PCA 的实际应用受到限制。ICA 与 PCA 一样,只能处理线性映射数据,但 ICA 求解的是最能使数据相互独立的方向,它将原始数据分解为若干独立分量的线性组合,它更适合处理非高斯分布的情况,即原始数据属于非高斯分布这种分解是唯一的情况。LDA 是有监督的特征提取方法,它与 PCA 一样,适合处理高斯分布数据。LDA 方法通过先计算类内散点矩阵  $S_w$  和类间散点矩阵  $S_b$ ,然后求矩阵  $S_w^{-1}S_b$  前  $n$  个特征值对应的特征向量,得到原始特征在新子空间的投影。LDA 方法让降维后的同类特征在新的子空间中尽量紧密,而不同类特征尽量分离。这种算法的缺点是需要满足数据样本呈正态分布且特征是随机独立的前提条件。多维尺度算法(MDS)与主成分分析都是用较少的维来表达观测数据,同属于无监督特征提取技术。但多维尺度算法利用的是低维空间中成对样本间的相似性与其在高维空间中的相似性尽量保持一致,缺点是计算量大且对噪声敏感<sup>[6]</sup>。

还有一些算法,例如:Wang<sup>[7]</sup>提出的随机投影算法,即当数据维数很大时,随机地将原始的高维数据映射到一个低维子空间,这种方法的计算效率很高;极小冗余-极大相关选择(mRMR)方法,其通过两个随机变量的互信息或熵等来考查特征之间、特征与类别之间的相关度,适用于离散数据,实际应用时需先将连续数据进行离散化。

### 2.2 非线性特征提取

在现实世界中,人们采集到的数据中有许多特征都呈非线性分布,如图像、视频、文本等,用线性模型来处理这类数据已无法捕捉它们的本质结构。研究者们越来越关注这个问题,并提出了一系列针对高维的非线性数据方法。它们主要分为两类,一类是基于核方法,这类方法的主要思想是将非线性数据通过核函数投影到维度更高的核空间,使数据线性可分。例如:Schölkopf 等<sup>[8]</sup>提出的核主成分分析 KPCA 就是以 PCA 为基础,用 Mercer 核将原空间的数据映射到高维空间,

使得在原空间线性不可分的数据样本在高维空间线性可分。Mika 等<sup>[9]</sup>利用同样的思想将 LDA 算法进行扩展,提出了核线性判别分析(KLDA)。此后,研究者们也提出了一些其他的核函数方法。例如:2006 年,Weinberg<sup>[10]</sup>提出了自大拉伸算法,其利用核矩阵的半正定约束关系来实现数据的优化。核函数的引入,使得这些核函数扩展方法的性能得到提高,但同时也一定程度上增大了样本维度,并且不同的核函数及对应参数的选择都会直接影响最后的结果,很多时候算法中的核函数和参数还需要靠经验来确定。

另一类是基于流形学习的方法。流形学习从一个新的角度来解决高维数据的特征提取问题,目前该方法已被应用在多个领域。流形学习方法假设数据点分布在一个嵌入高维空间的非线性低维流形上,它的目标是寻找一个转换矩阵,以将高维空间中的数据集投影到低维空间中,使得在高维空间中互为近邻的两点投影到低维空间后仍互为近邻。这种特征提取方法的计算效率较高,而且可获得全局最优解,是近年来特征提取技术的一个研究热点。2000 年,Science 上刊出较有影响的关于流形学习的文章提出了等距离映射算法(ISOMAP)<sup>[11]</sup>、局部线性嵌入算法(LLE)<sup>[12]</sup>等这让基于流形学习的非线性特征降维受到关注并成为了近年来的研究热点。

等距离映射算法(ISOMAP)<sup>[11]</sup>是 MDS<sup>[6]</sup>的推广。对于非线性数据,欧氏距离不能反映它们之间真实的位置关系,因此 ISOMAP 用测地距离替代了 MDS 算法中的欧氏距离,通过计算在低维空间中样本对之间的测地距离来保持数据点的内在关系。局部线性嵌入算法(LLE)<sup>[12]</sup>认为,如果一个数据集是从光滑流形中取样,那么与每个点相邻的点依然相邻而且也位于低维空间中。在低维空间中,LLE 利用其邻近点来线性重构一个样本点,并且重构权值保证样本点与领域点的重构误差最小,空间中的每个样本点都可以重构,在低维空间中样本点和近邻点的关系不变。除此之外,Belkin 等<sup>[13]</sup>提出了拉普拉斯特征谱算法(LE),他认为现实世界中相邻的样本在低维流形中应该仍然保持相邻,算法通过对点之间邻域关系无向图的操作来保持数据之间的近邻关系,图中拉普拉斯矩阵可替代流形中拉普拉斯 Beltrami 算子来实现高维向量向低维空间的嵌入。He 等提出的局部保持映射算法<sup>[14]</sup>先假设数据点分布在高维空间流形的前提下,然后利用高维数据局部线性的特点,通过一个转换矩阵将高维空间中的数据映射到低维空间中。海森特征谱方法(HE)<sup>[15]</sup>的提出主要是为了处理 ISOMAP 算法中数据采样不均匀的问题,Hessian 特征映射中使用 Hessian 算子替代 Laplacian 特征映射框架中的拉普拉斯算子,然后在 Hessian 算子的矩阵上运用特征值分解得到最终的低维坐标。多层感知器(MLP)<sup>[16]</sup>是一种前馈结构的神经网络,它的隐含层有奇数个神经元,可以将数据从高维空间映射到低维空间。由于网络中各层之间有大量连接,因此算法的收敛速度慢、效率低,而且容易陷入局部最优。

目前,已有越来越多的研究者关注流形学习。流形学习算法已经成为解决非线性分布数据的有效方法,但它仍然存在许多缺点,如对噪声敏感、泛化能力差等。国内外的研究者原算法框架的基础上做了改进,新方法虽然解决了目前流形学习中的一些问题,但还需进一步完善。

表1 对上述特征提取方法进行了总结。

表1 特征提取的主要方法

	方法	有/无监督
线性	主成分分析 PCA	无
	独立成分分析 ICA	无
	线性判别分析 LDA	有
	多维尺度算法 MDS	无
核方法	核主成分分析 KPCA	无
	核判别分析 KLDA	有
	最大差异伸展算法 MVU	无
非线性	等距离映射 ISOMAP	无
	局部线性嵌入 LLE	无
	拉普拉斯特征谱方法 LE	无
	局部保持映射 LPP	无
	海森特征谱方法 HE	无
	半正定嵌入算法 SDE	无

### 3 特征选择

特征选择是根据特征的相关性和冗余度从原始数据特征集中选出一些特征,使其满足特定指标的过程,选出的特征是原始特征的子集,因此也被称为特征子集选择。特征子集的产生过程<sup>[17]</sup>其实是一个搜索寻优的过程,可以分为5个阶段:搜索方向、搜索策略、评价准则、停止条件和验证结果。特征选择的过程从决定搜索起点和方向开始;然后根据不同的搜索策略进行子集搜索;特征子集产生后,用评价准则计算特征的重要性,然后进行选择,满足停止准则条件后停止特征选择过程;最后再用错误估计或者一些验证方法来评价特征子集用于分类是否有效。

特征选择过程中的搜索策略与评价规则至关重要。搜索策略是指特征选择算法用某种方法从特征空间中找到符合评价标准的特征子集,好的搜索策略能提供良好的全局搜索能力,可快速收敛到最优解附近,有较高的计算效率。按照特征子集形成过程的不同,将搜索策略分为3种:全局最优搜索、随机搜索和启发式搜索<sup>[18]</sup>。

#### 3.1 基于搜索策略的特征选择方法

##### 3.1.1 全局最优搜索策略的特征选择方法

全局最优搜索方法从原始特征集开始,逐一评价产生的每个子集,然后选出满足特定条件的特征子集,它能保证得到最优特征子集。例如,原始数据有 $d$ 维特征,则会产生 $2^d$ 个特征子集,依次对每一个子集进行评价,并从中选出一个满足特定指标的子集,该过程的时间复杂度为 $O(2^d)$ 。特别地,对于高维数据,其时间复杂度较高,在实际应用中很难应用。这类搜索方法主要有穷举法和分支定界法。

穷举法(Exhaustive)通过遍历,从原始数据的 $d$ 个特征中选出最优的 $m$ 个特征形成特征子集,这里共有 $C_d^m$ 种不同的特征组合。在实际应用中,穷举法的时间复杂度很高,属于NP-hard问题<sup>[19]</sup>,阻碍了其实际应用。分支定界法(Branch and Bound)的主要思想是,先定义一个满足单调性的评价准则函数,然后在原始特征集中用树搜索算法选择一个在给定的评价函数下最优的最小特征子集。由于使用不同的搜索方法可以提高搜索效率,因此与穷举法相比,分支定界法节省了计算时间,但对于高维数据,其运算时间仍然很长。分支定界法的关键是要求评价函数满足单调性,而实际应用中难以找到满足单调性的评价函数,而且需要提前给出子集特征的个数,因此该算法也不能被广泛应用。

全局搜索策略存在的主要问题是需要在搜索前先确定子集的数量;另外,处理高维数据时算法要多次运行,计算开销很大。

##### 3.1.2 随机搜索策略的特征选择方法

随机搜索策略从随机选择的特征开始,然后可用顺序搜索或双向搜索的方式进行搜索(如模拟退火算法),也可以用没有规则的运动的策略来搜索(如遗传算法、Las Vegas算法等)。根据分类的有效性对算法中的每个特征赋予权重,然后用定义的阈值来对特征进行评价,选出权重超过阈值的特征来训练分类器<sup>[17]</sup>。

遗传算法(GA)是借鉴进化论而提出的一种搜索方法,它在特征选择领域被广泛应用。算法首先随机产生一批特征子集,通过评价函数给这些特征子集进行评分,然后通过繁殖产生新一代的高评分特征子集。基于遗传算法的搜索方法不容易陷入局部极值,但搜索过程中会按照重要性对所有特征进行排序,计算开销较大。另外,将评分高的重要特征组合后形成的特征子集未必是最优特征子集。模拟退火算法(SA)是根据固体物质退火过程中系统温度缓慢地降低到零温度状态而提出的,它能接受在降温过程中能量可能会增加的情况,这个特点使得SA算法能跳出局部最优解,达到全局最优解。但这种算法的计算量较大,并且起始温度、退火速率和每个温度下的迭代次数都难以设定。

随机搜索算法易受随机因素的影响,具有较高的不确定性,因此每次实验结果都很难重现,只有当循环次数比较大时才能有好的结果。若数据有 $d$ 维特征,则它的时间复杂度仍为 $O(2^d)$ ,只是可以通过设置迭代次数使时间复杂度略小于 $O(2^d)$ 。另外,在随机搜索过程中设置不同的参数也对结果有不同的影响。

##### 3.1.3 启发式搜索的特征选择方法

启发式搜索算法也叫序贯优选法,它在搜索最优子集与计算复杂度之间做平衡。代表性算法有:广义序列前向搜索法(GSFS)、广义序列后向搜索法(GSBS)、浮动搜索方法等。

最早的序列前向选择方法(SFS)中,从空集开始,每次选择一个特征加入,直到特征函数最优为止。序列后向选择方法(SBS)特征子集从全集开始,每次选择一个特征剔除,直到特征函数最优。这两种算法每次只能增加(或减少)一个特征,直到得到最优子集。改进SFS算法后,得到广义序列前向搜索法(GSFS)。GSFS算法每次从未选入的特征中选择 $L$ 个特征加入到已选的特征子集中,直到这 $L$ 个特征加入后评价函数的值达到最大。对SBS算法进行改进后,得到广义序列后向搜索法(GSBS)。GSBS算法首先定义评价函数,然后每次删除对评价函数贡献最小的 $R$ 个特征,直到评价函数达到最优。双向搜索算法(BDS)同时使用序列前向选择算法和序列后向选择算法,从空集与全集同时开始搜索,直到两者搜索到相同的特征子集时停止。

比较常用的搜索算法还有增 $L$ 去 $R$ 选择算法和浮动搜索方法。增 $L$ 去 $R$ 选择算法将SFS和SBS结合起来,先使用SFS算法逐个选入 $L$ 个最佳特征,然后使用SBS算法再逐步剔除 $R$ 个最差特征,直到选出的特征个数达到要求为止。这种方法克服了前述算法存在的特征一旦选入就不能删除的缺点,但是每次增加和剔除的特征个数固定不变。浮动搜索方法针对增 $L$ 去 $R$ 选择算法中每次增加和剔除的特征个数不变的缺点,允许每次增加和剔除的特征个数可变。相比前述

的方法,该方法能更加迅速地接近全局最优<sup>[20]</sup>。

启发式搜索方法能对所有特征对算法规则的作用进行排序,对于  $d$  个特征,它的时间复杂度为  $O(d^2)$ ,可见这类方法的时间复杂度较低,可以快速收敛。另外,启发式搜索得到的特征子集比随机搜索和全局搜索得到的特征子集更小,算法的复杂度更低、效率更高;但它不能保证得到全局最优解。

### 3.2 基于特征子集评价准则的特征选择方法

对原始特征空间降维后的结果进行评价,判断所得到的特征维数以及所使用的特征是否对原分类最有利。评估方法在特征选择过程中非常重要,它对后续算法影响极大,是特征选择算法中的关键步骤。文献<sup>[21]</sup>根据评估方法来评价特征选择算法,将特征选择算法分为 Filter 模型、Wrapper 模型与 Embedded 模型。

#### 3.2.1 基于 Filter 评价的特征选择算法

Filter 评价策略的特征选择方法中的学习方法独立于具体的分类模型,如图 2 所示,它直接利用数据的统计特性来评价特征,并根据特征的重要程度来排序,进而去除得分较低的特征。

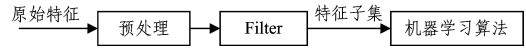


图 2 Filter 模型特征选择框图

文献<sup>[22]</sup>中将 Filter 模型特征选择的评估方法分为 4 类:距离度量、相关性度量、一致性度量和信息度量的评估方法。距离度量是一种常用的度量方法,有欧氏距离、马氏距离、平方距离等。一些典型的算法,如分支定界法和 Relief 系列算法等,都是基于距离度量的。文献<sup>[23]</sup>提出将核方法应用到距离度量中,有效地扩展了线性到非线性数据特征选择的能力。相关性度量用于统计相关系数,如 Pearson 相关系数、Fisher 分数、最小平方回归测度<sup>[19]</sup>等,它会得到特征与类之间的相互关系。一致性度量可以快速找出能够与完整特征集分类效果一致的最小特征子集,以去除冗余及噪声,但这种度量对噪声敏感且只适用于离散特征。信息度量是利用互信息、信息增益、关联度等来衡量特征的。这类度量方法中,度量函数要保证所选择的特征子集与类别的相关性最大,同时子集之间的相关性最小。表 2 对上述评价方法进行了总结。

表 2 Filter 模型特征选择算法的主要评价方法

度量方法	评价	评价方法	说明
距离度量	欧氏距离	$d = (\sum_{i=1}^n (x_i - y_i))^{\frac{1}{2}}$	距离度量主要对特征变量 $x$ 与 $y$ 之间的相异性进行测量;
	M 阶马氏距离	$d_M = (\sum_{i=1}^n (x_i - y_i)^m)^{\frac{1}{m}}$	同类特征中样本之间的距离应尽可能小,不同类特征样本之间的距离应尽可能大;
	S 阶 Minkowski 距离	$d_S = (\sum_{i=1}^s  x_i - y_j ^s)^{\frac{1}{s}}$	距离度量中准则函数应具有单调性,不能处理交叠的情况;
	Chebyshev 距离	$d = \max_i  x_i - y_i $	马氏距离不用考虑量纲。
相关性度量	Pearson 相关系数	$\rho(x, y) = \frac{\sum_i (x_i - \bar{x})(y_{ik} - \bar{y}_k)}{(\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2)^{\frac{1}{2}}}$	相关性度量评价是假设好的特征子集中包含的特征分类的相关度高,同类特征之间的相关度低;
	最小平方回归测度	$R^2 = \frac{\sum_k (\sum_i (x_i - \bar{x})(y_{ik} - \bar{y}_k))^2}{(\sum_i (x_i - \bar{x})^2 \sum_{i,k} (y_{ik} - \bar{y}_k)^2)}$	Pearson 系数用于衡量特征与类别的可分离性程度。
一致性度量	不一致率	找到与全部数据有同样能力的最小子集	单调快速,但只适用于离散值特征。
信息度量	信息增益	$IG(T) = H(C) - H(C T)$	互信息是非线性相关测度,可以处理离散特征,
	互信息	$MI(X Y) = H(X) - H(X Y)$	对连续变量需要进行离散化处理。

Filter 评价策略的特征选择方法选出的特征子集规模较大,可能存在评价结果与后续学习算法性能差异大的情况,但它可以快速地排除大量的不相关特征,可用作特征预筛选器。

#### 3.2.2 基于 Wrapper 评价特征选择算法

Wrapper 评价特征选择方法需要预先设定分类器学习模型,然后根据后续分类器的准确率对特征子集进行评价。Wrapper 评价特征选择方法的过程如图 3 所示,这类方法与所使用的分类器的关系很大,它将子集的选择视为一个搜索寻优问题,不同的分类器会生成不同的子集组合,用分类器评选验证集中的子集,并进行排序组合的比较。将贝叶斯分类器、最近邻方法、支持向量机<sup>[24-26]</sup>等理论方法与算法相结合来评价特征学习,根据最终的分类结果错误率来评价子集的好坏。Forman 对错误率的测度进行了汇总,常用的错误率测度有错误率、卡方、几率比、概率比等<sup>[27]</sup>。

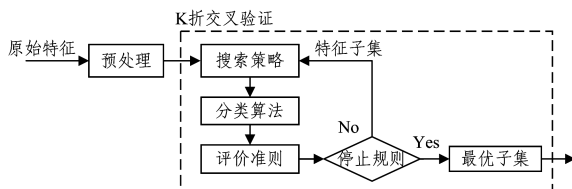


图 3 Wrapper 模型特征选择框图

基于 Wrapper 评价特征选择方法的优点是最后得到的特征子集的规模比 Filter 特征选择结果小,而且准确率高,但它会枚举出所有可能的特征组合,而且每种组合都会被用于训练模型及测试集的验证,因此它的计算速度慢,特别是在数据维度很大时会显著地增加计算的时间,而且泛化能力弱,不适合大数据集。

#### 3.2.3 基于 Embedded 模型的特征选择算法

Embedded 模型方法将特征选择作为学习算法的一部分,它同时解决了特征选择和分类或者聚类等问题,如图 4 所示。分类器依赖于选择的特征模型以及具体的学习算法<sup>[28]</sup>。这类特征选择方法主要有基于决策树模型的特征选择法、支持向量机(SVM)特征选择算法等。SVM 理论基于统计学理论和结构风险最小化原则,可以处理非线性分类问题,用最少的样本得到最小的分类误差,因此可以降低分类器的复杂程度。近年来,许多研究者都将特征选择的过程融入到向量机分类器的学习过程中。例如,Mangasarian 等用特征训练样本乘以一个与特征维度相同的矢量  $\sigma$ ,矢量中每个分量都由 0 或 1 构成。若分量为 0,表示对应的特征被删除;若分量为 1,表示对应的特征被保留。用这种方法将特征选择问题转化为求解表达式规划问题,在求解表达式解的同时也得到了最小的有效特征子集<sup>[29]</sup>。Wang 等<sup>[30]</sup>提出把 SVM 对多类特征分

类的过程融入特征选择算法中。

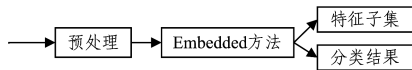


图4 Embedded模型特征选择框图

基于 Embedded 模型的特征选择方法同时具有 Filter 方法与 Wrapper 方法的优点,训练分类器与特征选择而且实现,不但提高了特征选择的速度,而且分类器的分类效果更好;其缺点是损失函数的选择以及缩放参数的确定都需要相关背景知识和经验。

#### 3.2.4 特征选择算法的其他评价策略

除了上述特征选择评价准则外,Leung 等<sup>[31]</sup>提出了混合的特征评价方法(Hybrid),它具有比 Filter 更高的准确率,比 Wrapper 的计算复杂度低,并且也不易过拟合,但它需要制定具体的分类器。文献<sup>[32-33]</sup>提出了基于集成的特征选择方法,该方法的目的是解决许多特征选择算法中的不稳定性问题,它用不同的采样策略,基于特征选择方法将多个样本所获得的结果进行合并,以形成更稳定的子集。基于集成方法的优点是对高维数据的降维更加灵活且不易过拟合,但在应用中存在一些待解决的问题。

## 4 特征降维技术的研究方向

特征降维处于模式识别的预处理阶段,受到人们的长期关注。过去几十年中已有大量的特征降维算法被深入研究,无论在理论上还是实际应用上都有较大的突破。随着大数据的出现,许多应用领域都出现了更多的数据信息,这些信息普遍具有维度高的特点,它们在给人们提供更丰富信息的同时也带来了许多的问题:人们处理的数据中出现了更复杂的数据对象以及新的数据类型、噪声,数据价值密度低等。这些问题使得传统的数据分析和挖掘任务等面临巨大的挑战。因此,针对某种具体应用的降维技术是现阶段模式识别领域的研究热点。下文总结了几个值得关注的研究方向。

### 4.1 混合模型特征降维算法

混合模型的特征选择方法结合了 Wrapper 模型与 Filter 模型的特点,它兼具了 Filter 模型的高效性和 Wrapper 模型的准确性,结合其互补的优势可得到更有效的特征子集。该算法可在算法的不同搜索阶段采用不同的评价标准,计算性能较好。由于支持向量机理论有扎实的基础,算法中多用 SVM 的分类结果作为评估标准。如:Wang 等<sup>[30]</sup>提出了基于 SVM 的混合特征选择算法,它先用 Filter 算法对原始特征集进行筛选,再用 Wrapper 算法从筛选后的特征子集中选择特征。该算法考虑了特征间的相关性,提高了计算效率。Li 等<sup>[34]</sup>也使用混合模型来选择特征子集,先用信息度量方法得到前  $k$  个重要的特征,再用基于 SVM 的分类器对预选的  $k$  个特征构成的特征子集的分类效果进行评分,从而得到最优特征子集。文献<sup>[35]</sup>中的混合模型算法结合了序列搜索算法,并将 ROC 曲线作为搜索策略,得到的特征子集明显提高了分类器的性能。

### 4.2 无监督的特征降维方法

监督的特征选择是最早、最普遍应用的方法,它利用数据的标签来评价。由于标记数据的获取需要先验知识,成本高昂且不易得到,而无标签数据获取更容易,因此近年来研究者们更多地采用半监督或无监督的特征选择方法。无监督的特征选择方法用的是无标签数据,它比监督和半监督的特征选

择方法更具挑战性,在生物数据分析等领域应用较多,是发现未知分类疾病的有效方法<sup>[36]</sup>。目前,研究者们提出的关于无监督的特征提取方法普遍存在一个问题:忽略了不同特征之间可能的关联,用这种方法得到的特征子集往往不是最优子集。在不同的应用领域,通过改进现有算法,采用无监督的特征选择方法得到较好的分类结果,是一个值得研究的方向。

### 4.3 基于集成的特征降维方法

基于集成的特征降维方法通过训练多个子集来生成个体学习器,并将个体学习器的结果进行合成,以达到提高集成个体差异度的目的。集成特征选择方法不再依赖某个特征选择的结果,不仅能解决特征选择算法中的不稳定问题,还能灵活地处理高维数据。实践证明,集成方法能提高模型的泛化性能和精度,成为近年来模式识别领域一个重要的研究方向<sup>[37]</sup>。相比于个体学习器,集成学习器有更高的学习精度,性能也得到提升。例如:Abeel 等<sup>[38]</sup>提出利用线性 SVM-RFE 的分类器实现集成的特征选择方法。近年来,对于集成学习在特征选择中的应用,研究者主要对学习过程中的个体特征选择、多任务学习等问题展开研究;另外,他们利用集成方法对有监督的特征选择方法进行改进,使其可有效用于无监督特征的选择<sup>[39]</sup>。

### 4.4 高维小样本数据的特征降维

在文本挖掘、计算机视觉、生物基因数据等领域,高维小样本数据相当普遍。这类数据的特征维数远大于样本个数,数据维度过高会产生大量的冗余信息,从而造成“维度灾难”;同时,样本数量不足又容易导致学习过拟合,传统的降维方法在处理高维小样本数据时往往失效,给现有的学习算法带来了巨大的挑战,因此高维小样本数据的特征选择成为了近年来的研究热点。研究者们针对高维小样本数据普遍存在的问题,分别提出了不同的解决办法。例如,为了降低计算的复杂度,避免过拟合,张靖等提出 K-split Lasso 来克服了 Lasso 方法选择特征个数较多、算法复杂的问题<sup>[40]</sup>;还有研究者提出用深度卷积网来训练海底生物图像,以解决高维小样本问题<sup>[41]</sup>。

### 4.5 深度学习的特征降维

2006年,Hinton 等<sup>[42]</sup>提出了神经网络特征降维方法,开始通过模拟人脑对图像多层抽象的过程来对图像数据特征进行表达,整个过程被认为是一种自动学习特征的方法。此后,神经网络学习成为了机器学习领域的研究热点,并已在语音、图像等多个领域都有广泛应用。深度学习的基础模块,如自编码器和受限玻尔兹曼机等,都能通过将隐层节点数调节到少于可视层的节点数来对原始数据进行降维;学习模型中选择某一神经层特征后,再进行最终目标模型的训练,可以提取出更加抽象的特征<sup>[43]</sup>,它的学习任务之一就是可以从原始数据中提取出适合的特征。目前,用深度学习法对高维数据进行降维时计算时间较长、开销大,因此该方向也有亟待解决的问题值得关注。

**结束语** 在信息时代,各类数据采集仪器的出现使得数据快速增长,大数据在带来更多信息的同时,也给研究者们带来了烦恼。要想合理地利用数据信息并避免“维度灾难”,在机器学习预处理阶段,数据特征降维的研究就显得尤为重要。

本文首先从特征提取和特征选择两方面对典型的特征降维技术进行阐述,并对比了其特点:在这两类方法中,特征提取是通过计算得到特征集合,通过变换得到的新特征无任何

物理意义,它们与类别之间的关系难以解释;特征选择方法多样,它选出的特征集合是原始特征集的子集,能很好地理解数据并有效地降低计算开销。然后,根据特征选择过程中,特征子集的形成过程和评价策略这两个关键环节对现有的特征选择方法进行总结,并对比了它们的特点。

降维技术从发展至今,已有许多成熟的方法。然而,随着大数据时代的到来,用传统的降维方法来处理“大数据”的效率较低甚至会失效,给机器学习带来了困难<sup>[2]</sup>。研究者们以现有的学习模型和研究理论框架为基础,针对不同的应用,对算法做了改进,提出了一些新的降维方法。本文根据各种方法的特点,归纳了新时期数据特征降维亟待解决的问题和值得研究的方向,它们对后续研究有着重要的意义。

### 参 考 文 献

- [1] SHEIK A. A Survey on Evolutionary Techniques for Feature Selection [C] // IEEE Conference on Emerging Devices and Smart Systems, Tiruchengode India; IEEE Press, 2017.
- [2] SAMINA K, TEHMINA K. A Survey of Feature Selection and Feature Extraction Techniques in Machine Learning [C] // Science and Information Conference, London; IEEE Press, 2014: 372-378.
- [3] JOLLIFFE I T. Principal component analysis [M]. Berlin: Springer-Verlag, 1986.
- [4] DUDA R O, HART P E, STORK D G. Pattern Classification (2nd Edition) [M] // En Broeck the Statistical Mechanics of Learning Rsity, 2000: 32-39.
- [5] COMON P. Independent component analysis, a new concept [J]. Signal Processing, 1994, 36(3): 287-314.
- [6] BRONSTEIN A M, BRONSTEIN M M, KIMMEL R. Generalized multidimensional scaling: a framework for isometry-invariant partial surface matching [J]. Proceedings of the National Academy of Sciences of the United States of America, 2006, 103(5): 1168-1172.
- [7] WANG J Y. Geometric structure of high-dimensional data and dimensionality reduction [M]. New York; Springer Heidelberg, 2011: 131-147.
- [8] SCHÖLKOPF B, SMOLA A, MULLER K R. Nonlinear Component Analysis as a Kernel Eigenvalue Problem [J]. Neural Computation, 1998, 10(5): 1299-1319.
- [9] MIKA S, RÄTSCHE E, WESTON J, et al. Fisher Discriminant Analysis with Kernels [C] // Proceedings of IEEE Workshop Neural Networks for Signal Processing, 1999: 41-48.
- [10] WEINBERGER K Q, SAUL L K. Unsupervised learning of image manifolds by semidefinite programming [J]. International Journal of Computer Vision, 2006, 70(1): 77-90.
- [11] TENENBAUM J B, SILVA V, UNGFORD J C. A global geometric framework for nonlinear dimensionality reduction [J]. Science, 2000, 290(12): 2319-2323.
- [12] ROWEIS S T, SAUL L K. Nonlinear dimensionality reduction by locally linear embedding [J]. Science, 2000, 290(5500): 2323-2326.
- [13] BELKIN M. Problems of learning on manifolds [D]. Chicago: The University of Chicago, 2003.
- [14] HE X F, NIYOGI P. Locality preserving projections [C] // Advances in Neural Information Processing Systems 16. Vancouver, Canada; MIT Press, 2003: 153.
- [15] DONOHO D L, GRIMES C. Hessian Eigenmaps: New Locally Linear Embedding Techniques for High-dimensional Data [J]. Proceedings of the National Academy of Sciences of the United States of America, 2003, 100(10): 5591-5596.
- [16] MOALLEN P, AYOUGH S A. Removing potential flat spots on error surface of multilayer perceptron (MLP) neural networks [J]. International Journal of Computer Mathematics, 2011, 88(1/3): 21-36.
- [17] JUNCHIN A, ANDRI M. Supervised, Unsupervised, and Semi-Supervised Feature Selection: A Review on Gene Selection [J]. Transactions on Computational Biology and Bioinformatics, 2016, 13(5): 971-989.
- [18] SUN Z H, GEORGE B, RONALD M. Object detection using feature subset selection [J]. Pattern Recognition, 2004, 37(11): 2165-2176.
- [19] CAI Z Y, YU J G, LI X P, et al. Feature selection algorithm based on kernel distance measure [J]. Pattern Recognition and Artificial Intelligence, 2010, 23(2): 235-240.
- [20] PUDIL P, NOVOCICOVA J, KITTLER J. Floating Search Methods in Feature Selection [J]. Pattern Recognition Letters, 1994, 15(11): 1119-1125.
- [21] LIU H, YU L. Toward integrating feature selection algorithms for classification and clustering [J]. IEEE Transactions on Knowledge and Data Engineering, 2005, 17(4): 491-502.
- [22] KOLLER D, SAHAMI M. Toward optimal feature selection [C] // Thirteenth International Conference on International Conference on Machine Learning, Morgan Kaufmann Publishers Inc., 1996: 284-292.
- [23] MITRA P, MURTHY C A, SANKAR K P. Unsupervised feature selection using feature similarity [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002, 24(3): 301-312.
- [24] GUYON I, WESTON J, BARNHILL S, et al. Gene selection for cancer classification using support vector machines [J]. Machine Learning, 2002, 46(1): 389-422.
- [25] YANG J B, ONG C J. Feature selection for support vector regression using probabilistic prediction [C] // 16 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2010: 343-352.
- [26] SHEN K Q, CHONG C J, LI X P, et al. Feature selection via sensitivity analysis of SVM probabilistic outputs [J]. Machine Learning, 2008, 70(1): 1-20.
- [27] FORMAN G. An extensive empirical study of feature selection metrics for text classification [J]. Journal of Machine Learning Research, 2003, 3: 1289-1305.
- [28] NG A Y. Feature selection, L1 vs. L2 regularization, and rotational invariance [C] // Proceedings of the Twenty-first International Conference on Machine Learning, New York; ACM, 2004: 78.
- [29] MANGASARIAN O L, WILD E W. Feature Selection for Nonlinear Kernel Support Vector Machines [C] // Seventh IEEE International Conference on Data Mining-workshops, 2007: 231-236.
- [30] WANG L F, SHEN X T. Multi-category support vector machines, feature selection and solution path [J]. Statistica Sinica, 2006, 16(2): 617-633.
- [31] LEUNG Y, HUNG Y. A multiple-filter-multiple-wrapper approach to gene selection and microarray data classification [J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2010, 7(1): 108-117.
- [32] LAZAR C, TAMINAU J, MEGANCK S, et al. A survey on filter techniques for feature selection in gene expression microarray analysis [J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2012, 9(4): 1106-1119.
- [33] SHEN Q, DIAO R, SU P. Feature Selection Ensemble [C] // Turing, 2012: 289-306.

量训练构建图像的数字字典,对大量未标记的遥感图像数据分类。

对于包含海量信息的遥感图像,如何充分挖掘信息以更贴合遥感图像分类要求成为了研究的重点。遥感图像涵盖的纹理特征、光谱特征和空间特征都可以单独作为图像分类的依据,然而图像的这些特征信息在图像分类时尚未充分利用。对提取的这 3 类特征进行多角度充分利用,并将其共同作为分类依据,结合深度神经网络的训练模型,来提高遥感图像的分类精度,成为了该领域目前的研究热点。

**结束语** 本文首先介绍了遥感图像分类的相关概念,分析遥感图像分类目前存在的问题,并对神经网络的历史、原理进行简要介绍,探讨了几种神经网络的原理结构;然后分别对遥感图像分类的研究现状和神经网络对遥感图像分类的研究现状做出阐述;最后总结了神经网络对遥感图像分类的发展趋势。

### 参 考 文 献

- [1] HINTON G E, SALAKHUTDINOV R R. Reducing the dimensionality of data with neural networks[J]. *Science*, 2006, 313(5786):504-507.
  - [2] 庞荣. 深度神经网络算法研究及应用[D]. 成都:西南交通大学, 2016.
  - [3] HINTON G E, OSINDERO S, TEH Y W. A fast learning algorithm for deep belief nets[J]. *Neural Computation*, 2006, 18(7): 1527-1554.
  - [4] 吴正文. 卷积神经网络在图像分类中的应用研究[D]. 成都:电子科技大学, 2015.
  - [5] LIPPMANN R P. Pattern Classification Using Neural Networks [J]. *IEEE Communications Magazine*, 2002, 27(11):47-64
  - [6] 卢柳叶,张青峰,李光录. 基于 BP 神经网络的遥感影像分类研究[J]. *测绘科学*, 2012, 37(6): 140-143.
  - [7] 卜晓波,龚珍,黎华. 基于遗传算法改进 BP 神经网络的遥感影像分类研究[J]. *安徽农业科学*, 2013, 41(33): 13056-13058, 13079.
  - [8] 胡永森,王力,吴良才,等. 加权变异粒子群 BP 神经网络在遥感影像分类中的应用[J]. *地理空间信息*, 2016, 14(12):37-40.
  - [9] 谭秀辉. 自组织神经网络在信息处理中的应用研究[D]. 太原:中北大学, 2015.
  - [10] 杜华强,范文义. Matlab 自组织神经网络在遥感图像分类中的应用[J]. *东北林业大学学报*, 2003(4):51-53.
  - [11] 李石华,金宝轩. 基于 Matlab 的自组织神经网络在地形复杂区遥感图像分类中的应用研究[C]//第二届“测绘科学前沿技术论坛”论文精选, 2010:4.
  - [12] 尹汪宏,李朝峰,张俊本,等. 基于混合核函数的自组织神经网络遥感图像分类[J]. *计算机工程与设计*, 2009, 30(2):388-391.
  - [13] 任军号,吉沛琦,耿跃. SOM 神经网络改进及在遥感图像分类中的应用[J]. *计算机应用研究*, 2011, 28(3):1170-1172,1182.
  - [14] 瞿继双,瞿松柏,王自杰. 基于特征的模糊神经网络遥感图像目标分类识别[J]. *遥感学报*, 2009, 13(1):67-74.
  - [15] 崔曦. 神经网络及模糊算法的遥感数据分类研究[D]. 西安:西安科技大学, 2008.
  - [16] 张强. 基于模糊神经网络的遥感影像分类研究[D]. 昆明:昆明理工大学, 2006.
  - [17] EINEN D, ROLFE J, FERGUS R. Understanding deep architectures using a recursive convolutional network[J]. *arXiv:1312.1847v2*, 2014.
  - [18] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks or larger-scale image recognition[J]. *arXiv:1409.1556*, Sept. 4, 2014.
  - [19] 曹林林,李海涛,韩颜顺,等. 卷积神经网络在高分遥感影像分类中的应用[J]. *测绘科学*, 2016, 41(9):170-175.
  - [20] 邢晨. 基于深度学习的高光谱遥感图像分类[D]. 武汉:中国地质大学, 2016.
  - [21] 王巧玉. 基于深度学习的高光谱遥感图像分类[D]. 厦门:华侨大学, 2016.
  - [22] 刘大伟,韩玲,韩晓勇. 基于深度学习的高分辨率遥感影像分类研究[J]. *光学学报*, 2016, 36(4):306-314.
  - [23] 黄鸿,何凯,郑新磊,等. 基于深度学习的高光谱图像空-谱联合特征提取[J]. *激光与光电子学进展*, 2017, 54(10):180-188.
  - [24] 付秀丽,黎玲萍,毛克彪,等. 基于卷积神经网络模型的遥感图像分类[J]. *高技术通讯*, 2017, 27(3):203-212.
  - [25] 张日升,张燕琴. 基于深度学习的高分辨率遥感图像识别与分类研究[J]. *信息通信*, 2017(1):110-111.
  - [26] LOPUHHIN K. Full pipeline demo: poly->pixels->ML->poly(Version4. 0) [EB/OL]. <https://www.kaggle.com/lopuhin/full-pipeline-demo-poly-pixels-ml-poly>.
  - [27] ZFTurbo. 0. 51276 Public LB Solution(Version1. 0) [EB/OL]. <https://www.kaggle.com/c/dstl-satellite-imagery-feature-detection/discussion/29829>
  - [28] ChrisCC. Object-based solution for DSTL (Version1. 0) [EB/OL]. <https://www.kaggle.com/chriscc/object-based-solution-for-dstl>.
- (上接第 21 页)
- [34] LI G Z, YANG J Y. Feature selection for ensemble learning and its application[M]// *Machine Learning in Bioinformatics*. 2008: 135-155.
  - [35] PENG Y H, WU Z Q, JIANG J M. A novel feature selection approach for biomedical data classification [J]. *Journal of Biomedical Informatics*, 2010, 43(1):15-23.
  - [36] CHIN A J, MIRZAL A, et al. Supervised Unsupervised, and Semi-Supervised Feature Selection: A Review on Gene Selection [J]. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2016, 13(5):971-989.
  - [37] OPITZ D W. Feature Selection for Ensembles[C]// *Proceedings of National Conference on Artificial Intelligence*. Orlando, FL, 1999:379-384.
  - [38] ABEEL T, HELLEPUTTE T, VAN D P Y, et al. Robust biomarker identification for cancer diagnosis with ensemble feature selection methods [J]. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2010, 26(3):392-398.
  - [39] WONG H S, ZHANG S, SHEN Y, et al. A New Unsupervised Feature Ranking Method for Gene Expression Data Based on Consensus Affinity[J]. *IEEE/ACM Transactions on Computational Biology & Bioinformatics*, 2012, 9(4):1257-1263.
  - [40] 张靖,胡学钢,张玉红,等. K-split Lasso:有效的肿瘤特征基因选择方法[J]. *计算机科学与探索*, 2012, 6(12):1136-1143.
  - [41] JIN L L, LIANG H. Deep Learning for Underwater Image Recognition in Small Sample Size Situations [C]// *IEEE Conference on Oceans*. Aberdeen UK: IEEE Press, 2017.
  - [42] HINTON G. Reducing the Dimensionality of Data with Neural Networks [J]. *Science*, 2016, 313(5786):504-507.
  - [43] 孙志远,鲁成祥,史忠植,等. 深度学习研究与进展[J]. *计算机科学*, 2016, 43(2):1-8.