

空间多维经济统计数据的降维方法 ——以四川省经济统计数据为例

董承玮^{1,2}, 芮小平^{1,3}, 邓 羽^{4,5,6}, 关兴良^{4,5}

(1. 中国科学院研究生院资源与环境学院, 北京 100049; 2. 北京市测绘设计研究院, 北京 100038;
3. 中国科学院生态环境研究中心, 北京 100085; 4. 中国科学院地理科学与资源研究所, 北京 100101;
5. 中国科学院研究生院, 北京 100039; 6. 哈佛大学, 美国坎布里奇 02138)

摘要: 经济统计信息往往包含多维属性, 需要采用降维方法将多维信息转换到三维以内的空间来实现多维信息可视化, 这有助于研究其内在空间分布规律。在评价线性方法 (PCA)、非线性方法 (NLM 和 SOFM), 以及监督分类方法 (SVM) 等四种降维方法的基础上, 以 2007 年四川省区县尺度为研究单元, 运用不同分类方法针对区县社会经济发展现状进行聚类 (分类) 处理, 并对成果的差异性展开了深入讨论, 主要结论如下: PCA 虽然能在整体上揭示经济发展趋势, 但结果与实际情况差异较大; NLM 能很好地展现出四川经济发展的区域态势和核心区域, 准确反映了四川经济发展现状; SOFM 的分类结果与发展现状较吻合, 但局部地区存在一定的错分情况, 且不能进行类内目标的比较; SVM 是监督分类, 需要已知样本来训练分类过程, 在样本的选择上存在较大的主观性, 且最优参数的搜索过程较为复杂。本文对几种降维方法的比较, 并在经济统计领域中的应用, 可以为相关的空间多维信息降维研究提供参考。

关键词: 降维; 多维可视化; 经济统计数据; 四川

文章编号: 1000-0585(2012)08-1411-11

1 引言

随着科学计算、工程测量的高速发展, 各个领域的信息积累量呈爆炸式增长。由于人类认知能力的固有局限性, 面对海量的多维信息和数据, 其信息认知和提取面临前所未有的挑战。因此, 在知识发现、信息认知和决策过程中, 多维信息可视化技术作为有效的抽象信息表达工具得到普遍运用, 辅助知识工作者理解和分析多维数据集的内在结构。降维方法^[1]可分为线性方法和非线性方法: 线性方法最常用的是主成分分析法^[2, 3]和多维尺度变换^[4, 5]; 而非线性方法有核主成分分析法^[6]、非线性映射^[7, 8]、神经网络^[9, 10]等。对于非线性结构的高维数据, 线性的降维方法不能准确地分析和提取其内在的结构; 而非线性方法的数学理论基础不同, 各种方法的分析结果也各不相同。

在对经济统计数据的研究中, 学者们较常使用的方法是 PCA 等线性方法, 然而这些

收稿日期: 2011-07-15; 修订日期: 2012-02-25

基金项目: 国家自然科学基金项目 (40901191)

作者简介: 董承玮 (1984-), 男, 湖南衡阳人, 硕士, 研究方向为三维 GIS、多维空间信息可视化。

E-mail: dongchengwei08@mails.gucas.ac.cn

通讯作者: 芮小平 (1975-), 男, 江苏苏州人, 副教授, 主要从事地理信息系统理论与应用方面的研究。

E-mail: ruixp@gucas.ac.cn

方法在分析非线性多维数据的过程中存在一定的局限性,而非线性方法的分析结果也各不相同。为揭示各种不同方法在分析多维数据时的优劣,本文实现了线性方法 PCA、非线性方法 NLM 和 SOFM,以及一种监督分类方法 SVM,对四川 2007 年四川统计年鉴数据进行分析,并以四川经济发展现状与规划为依据,对这几种结果进行分析,以说明各种方法的分类结果之间的差异及其优劣程度。

2 降维方法简介

2.1 主成分分析 (Principal Component Analysis, PCA)

PCA^[2, 3]是使用最广泛的线性降维方法之一。它的概念简单,实现算法高效,是把原来多个变量划为少数几个综合指标的一种统计分析方法。PCA 将方差的大小作为衡量信息量多少的标准,认为方差越大,提供的信息越多,反之提供的信息就越少。其原理是设法将原来变量重新组合成一组新的互相无关的几个综合变量,同时根据实际需要从中可以取出几个较少的总和变量尽可能多地反映原来变量的信息。

2.2 非线性映射 (Nonlinear Mapping, NLM)

MDS^[5, 6]是一种将目标对之间的相似性(或相异性)表示为低维空间中的距离的方法,它是一种应用广泛的线性降维算法,在图像处理、计算机视觉等方面有着广泛的应用。如对某多维数据集, MDS 将每条记录视为一个点,然后表达在低维空间中(二维或三维),并且点之间的相似度越大,距离就越近。借助于这种数据记录之间关系的图形表示方法,就可以观察数据和可视化研究数据集的结构,达到简化数据、揭示数据潜在规律的目的。原始和投影后的距离矩阵的一致性用一个误差函数来衡量,称为代价函数(Cost Function),如 Kruskal 压力函数^[6]:

$$s = \sqrt{\frac{\sum_{i < j} (\delta_{ij} - d_{ij})^2}{\sum_{i < j} \delta_{ij}^2}} \quad (1)$$

基于相同的理论, Sammon 提出了非线性映射^[7] (Nonlinear Mapping, NLM),如同 MDS, NLM 试图在二维或三维空间中保持高维数据的局部几何关系。其误差函数^[7]为:

$$E = \frac{\sum_{i < j}^k [d_{ij} - \delta_{ij}]^2 / d_{ij}}{\sum_{i < j}^k d_{ij}} \quad (2)$$

与经典 MDS 算法相比, NLM 误差函数中 $(1/d_{ij})$ 的存在, 对大相异度赋予更小的权重, 这减弱了大相异度在误差函数中的贡献, 在最小化目标函数的过程中, 能更好地对低相异度进行保持^[8]。如此数据可以平稳地映射到二维空间中, 这种非线性迭代过程能更好地对数据进行可视化, 更适合实际数据的降维映射^[7]。

2.3 自组织特征映射 (Self-Organizing Feature Map, SOFM)

自组织映射 (Self-Organizing Feature Map, 简称 SOFM)^[11~13], 是由芬兰赫尔辛基大学神经网络专家 Teuvo Kohonen 教授在 1981 年提出的竞争式神经网络, 它模拟大脑神经系统自组织特征映射的功能, 在训练中能无监督地进行自组织学习。它是一种聚类和高维可视化的无监督学习算法。SOFM 算法以其所具有的无监督学习、可视化、拓扑结构保持以及概率保持等特性, 广泛应用于聚类分析、图像处理、语音识别等信息处理领域。

SOFM 的网络拓扑包括输入层和输出层: 输入层由 N 个输入神经元组成; 输出层也称为竞争层, 由 M 个输出神经元组成, 这 M 个单元位于低维(通常为一维或者二维)规

则网格中，其基本单元的形状可以是四边形等基本几何图形，每个神经元代表了一个类别。网络是全连接的，即每个输入结点都同所有的输出结点相连接，连接的权值初始值可随机设定。

SOFM 网络聚类的基本思想是通过网络训练，把相类似的输入数据映射到同一个或者邻近的输出结点上，从而实现对输入数据的聚类和高维数据的低维映射。

2.4 支持向量机 (Support Vector Machine, SVM)

支持向量机 (Support Vector Machine, SVM)^[14~17]是 Corinna Cortes 和 Vapnik 等于 1995 年首先提出的，它在解决小样本、非线性及高维模式识别中表现出许多特有的优势^[9]，并能够推广应用到函数拟合等其他机器学习问题中。现今已经在许多领域如生物信息学，文本和手写识别等，都取得了成功的应用。

支持向量机方法是建立在统计学习理论的 VC 维 (Vapnik-Chervonenkis Dimension)^[17]理论和结构风险最小原理基础上的，它根据有限的样本信息在模型的复杂性（即对特定训练样本的学习精度）和学习能力（即无错误地识别任意样本的能力）之间寻求最佳折衷，以期获得最好的推广能力。支持向量机可以自动寻找对分类有较好区分能力的支持向量，由此构成的分类器可以最大化类与类之间的间隔。

SVM 是从线性可分情况下的最优分类面发展而来的，基本思想是由两类线性可分问题发展而来的。在线性可分数据的分类情况下，分类超平面是指能把样本正确分类的超平面，而分类间隔是指平行于分类超平面且过两类中离分类超平面最近的样本的两个超平面的距离，而这些距离最近的样本就是所谓的支持向量 (Support Vector, SV)，它们决定了最优分类超平面，使得属于两个不同类别的数据点不仅能被正确地分开，而且他们之间的间隔也最大。SVM 的实现过程就是首先寻找对分类有较好区分能力的支持向量，然后求解最优分类超平面。

低维空间中的向量集通常不是线性可分的，其解决的方法是将它们非线性映射到高维空间，但同时增加了计算复杂度，而利用核函数^[18, 19]可以巧妙地解决了这个问题，只要选用适当的核函数，就可以得到高维空间的分类函数。在 SVM 理论中，采用不同的核函数将形成不同的 SVM 算法，常用的核函数^[20]有线性核函数、多项式核函数、径向基核函数 (Radial Basis Function, RBF) 和 Sigmoid 核函数。实验证明径向基核函数相比其它核函数不仅具有较少的参数还具有良好的性能^[21]。

在 RBF 核函数中，惩罚系数 C ^[22, 23]表示由训练样本产生的经验风险对模型影响程度，它控制对错分样本惩罚的程度。 C 的取值大，对经验误差的惩罚大，经验风险较大，导致“过学习”； C 的取值小表示对经验误差的惩罚小，训练误差较大，导致“欠学习”。在调整过程中，训练误差随 C 的增大而单调下降，当 C 增至一定值后，训练误差趋于稳定，因此在能够解决问题情况下，应减小 C 值减小模型经验风险。RBF 核参数 γ ^[23, 24]的取值对 SVM 性能优劣起着关键作用，当 γ 取值很小时，SVM 会对训练集造成过学习现象，而太大则产生欠学习现象。目前对参数 (C, γ) 的选择主要有双线性搜索法^[23, 25]，网格搜索法^[23, 25]，梯度下降法^[24]、遗传算法^[26]等。

从本质上看，SVM 是一种监督分类方法，在对数据进行分类时，必须先要有一个已经正确分类的小样本集。由于在本文利用的经济统计数据中没有已知的正确分类样本集，作者利用上述 PCA 等方法得出的结果选取小部分样本来代替已知样本集，然后导入 SVM 模型中，从而得到分类结果。

3 研究区概况与数据来源

3.1 研究区概况

四川省位于我国西南地区、长江上游,东经 $97^{\circ}21' \sim 108^{\circ}31'$,北纬 $26^{\circ}03' \sim 34^{\circ}19'$ 。其中,成都市作为四川省的政治、经济、金融中心,其经济发展受政策、商业投资^[27]等影响,发展良好;广大东部和东南地区以平地、丘陵、平坝为主,交通发达,区位优势明显,是四川省内城镇最为密集的地区,也是西南地区经济发展水平最高的地域;川南地势起伏较大,海拔也高,地貌类型复杂多样,但资源非常丰富,近年来大力开发旅游资源,经济较为发达;川西和川北是高原山地,基础设施建设比较滞后,产业基础薄弱,东西部社会经济发展差距较明显。

目前,四川已逐步形成了五大经济区域^[27~31]:成都平原及其周边地区、川东南地区、川东及川东北地区、攀西地区,以及川西高原地区。从经济规模来看,成都平原及其周边地区占四川总量的 60% 左右,为各地区之首;其次为川东及川东北地区,占总量的 20% 左右;川东南占 15% 左右,攀西地区占 5% 左右;而川西高原地区不到 2%。而从人均 GDP 看,成都平原及其周边地区最高,但也只与全国平均水平相当,而其他区域都大大低于平均水平。

上述五大经济区域中,形成了四个经济核心区域:成都平原地区,包括成都、德阳、绵阳、乐山和眉山 5 个地级及地级以上城市,还包括崇州、江油等 9 个县级市,这是全省的经济核心地区;川南地区,包括内江,自贡,宜宾和泸州市和隆昌县,形成川南环线;川东北的遂宁—南充—广安地区;攀枝花—西昌轴线地区。这四个地区中成都平原地区是最发达的地区,而遂宁—南充—广安地区经济实力较弱,只是经济相对密集的区域。

3.2 数据说明

以 2007 年四川统计年鉴数据为例,利用上述方法对其进行分析。在行政单元为区县的经济统计数据中,大量属性维度的数据统计不完整。基于降维过程的维度应尽量最大化及其可获得性考虑,本文选择统计年鉴中最能反映地区经济发展情况的 18 个属性,包括国内生产总值(第一、第二、工业、第三产业和人均生产总值)、民营经济生产情况(第一、第二、工业、第三产业和人均民营经济增加值)、从业情况(从业人员、职工人数、人均工资)、地方财政(财政收入和支出)、农林牧渔总产值、社会消费品零售总额、全社会固定资产投资。

4 结果分析

4.1 PCA 的分类结果

基于 PCA,并设置方差舍弃阈值为 90%,对四川经济数据进行分析,并对结果进行分类,结果如图 1 所示。该分类结果整体上能体现出四川区域经济的发展状况,呈现出以成都市及西南部攀枝花市为核心的第一等级格局。成都作为省会城市,拥有较好的经济发展基础,自 2007 年 6 月成为国家综合改革配套试验区之后,成都市更是加快了产业结构调整步伐,新一轮的投资和创新拉动必然带动经济的快速发展。成都周边和东北部交通便利,并且受到省会经济发展的辐射作用,各区县也发展良好;川东南和川东北存在各区域的经济发展的核心;而广大西部的交通不便和基础设施薄弱造成发展的滞后。PCA 分类结果未能完全体现出成都周边区县的经济水平,也未能展现出川南环线区域,同时将川

东北大部分区县划为第四等级，而实际上东北各县经济较西北各区县发达。

4.2 NLM 的分类结果

基于 NLM 降维算法，并将统计数据集降至一维，分类结果如图 2 所示。基于四川的经济发展现状，该结果能反映真实的经济发展情况：该分类结果将成都大部分区、攀枝花东区和宜宾市的翠屏区划分为第一等级，将成都周边各县和东北部一些区县划分为第二等级，将东北部剩余区县划分为第三等级，而西部和东南部广大地区被划为第四等级，这从整体上充分体现了四川五大经济区域和四个核心发展区域的空间分布格局，但第二和第三等级之间、第三和第四等级之间的分类细节无法得到证实。

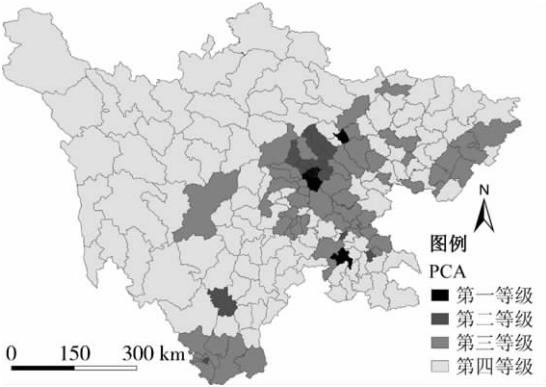


图 1 PCA 方法的分类结果
Fig. 1 The classification result of PCA

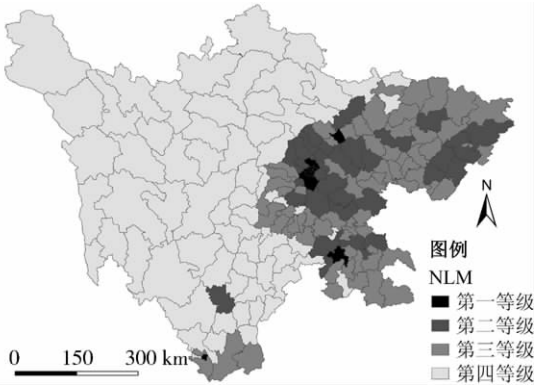


图 2 NLM 一维分类结果
Fig. 2 The classification result of NLM

4.3 SOFM 的分类结果

图 3 为 SOFM 的分类结果，与 NLM 的结果类似，SOFM 的分类结果从整体上体现了四川经济发展格局，但在四川东北部，除遂宁—南充—广安地区外，周边大部分区县也被划至第二类，未能体现出东北经济核心发展格局；其等级之间的分类细节也无法得到验证。

4.4 SVM 分类结果分析

从 PCA 结果的每类中心处提取 3 个样本作为已知小样本集导入 SVM 模型中，并考虑到 RBF 核函数的良好性能，利用网格搜索法对多个 (C, γ) 值对进行实验。结果

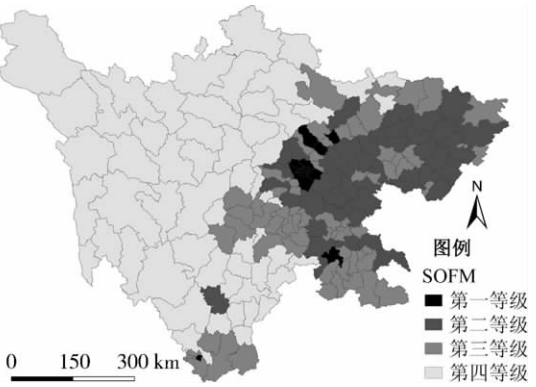


图 3 SOFM 分类结果
Fig. 3 The classification result of SOFM

表明，当 γ 较大时，结果逐渐出现欠学习问题，东北大部分区县被划至第四等级（图 4a），而第一等级范围也逐渐扩大（图 4b）。而当 γ 在一定合理范围之内（0.01~0.1），改变 C 值时，分类结果较稳定，各分类结果的不同主要体现在第三和第四等级的分类上，同时实验证明 $\gamma=0.01$ 时结果更为稳定。通过参数搜索实验，当 γ 取值在 0.01~0.1 区间（图 4c~4f）， C 值在 1~100 区间时，结果较为稳定，同时在参数搜索范围内更能反映发

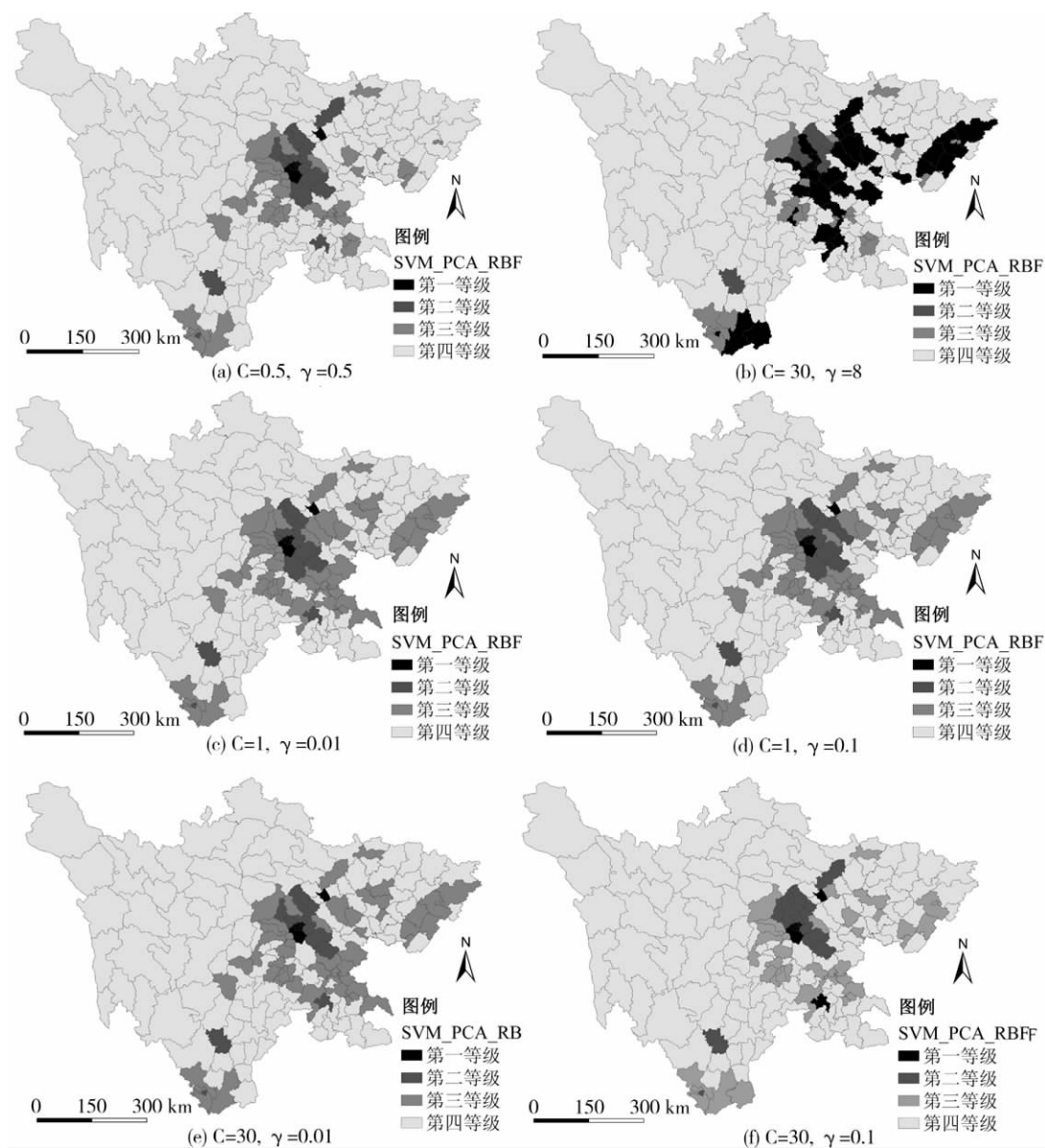


图 4 RBF 核函数 (PCA) 的分类结果

Fig 4 The result of RBF-SVM (PCA)

展现状, 是比较正确的结果。

从 NLM 结果的每类中心处提取 3 个样本作为已知小样本集导入 SVM 模型中, 同时考虑到 RBF 核函数的良好性能, 作者利用网格搜索法对多个 (C, γ) 值对进行实验。当 (C, γ) 值对很小时, 东北区县的分类结果不能体现经济发展空间发展格局 (图 5a), 而随着 C 的增大, 分类结果逐渐展现出四川经济发展核心区 (图 5b); 而当 γ 过大时, 分类结果完全出现错误 (图 5c); 当 γ 在 $[0.5, 2]$ 区间时, C 值在 $[10, 100]$ 区间时, 分类结果都高度相似, 且能体现总体发展格局 (图 5d~5f)。

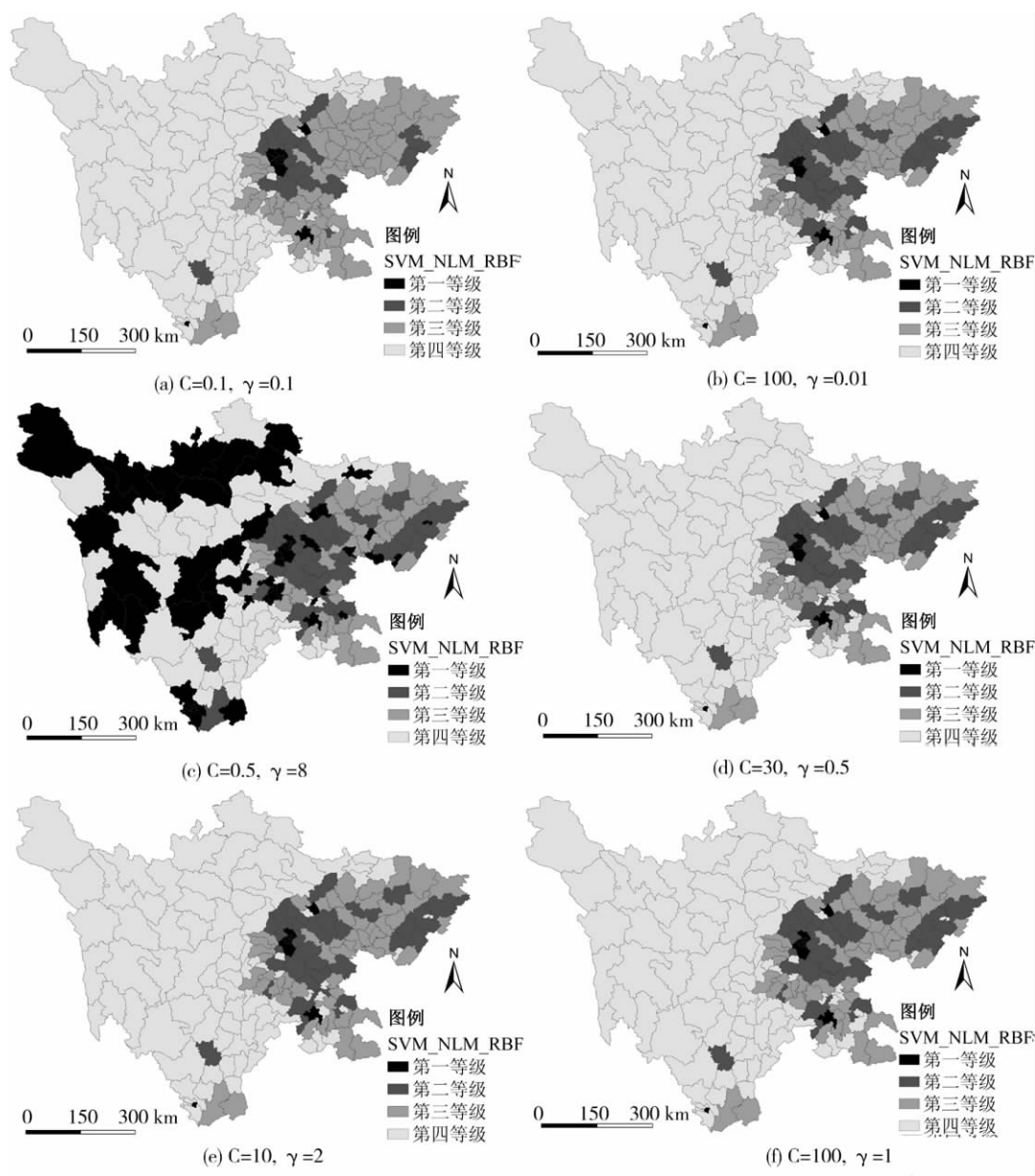


图 5 RBF 核函数 (NLM) 的分类结果
Fig 5 The result of RBF-SVM (NLM)

4.5 各方法分类结果的比较

表 1 总结了各方法分类结果的差异，分析结论如下：

(1) 从表 1 的等级分布来看，基于 PCA 的分类结果与其他方法的结果有较大不同，主要表现在第二、第三等级和第四等级的差异上，表现为只有成都周边的少数区县被分为第二等级，而其周边大部分区县和三大经济核心区域被划至第三等级，而大部分区县属于第四等级，未能准确体现出四川经济发展格局，这是因为线性算法不能完全挖掘非线性数据内在结构；但 PCA 能在降维之后求得各个目标的分值，通过分值的排序可以了解各县

区社会经济发展水平的整体次序情况。

表 1 各种方法的等级分布比较

Tah 1 The comparison of classification results of these methods

方法	第一等级	第二等级	第三等级	第四等级
PCA	金牛区、成华区、青羊区、武侯区、锦江区、双流县、涪城区(绵阳市)、翠屏区(宜宾市) 8 个	旌阳区(德阳市)、都江堰市、温江区、东区(攀枝花市)、西昌市等 14 个	广安市等东北部、眉山市等成都周边、仁和区等川南共 49 个区县	阿坝州等广大川西和川北高原地区,西昌周围的多数南部区县,以及川东北部分区县,共 110 个区县
PCA-SVM (C=30, γ=0.01)	上述除去翠屏区的 7 个	旌阳区、都江堰市、东区等 15 个	分布同上,共 51 个区县	分布同上,共 108 个区县
NLM	PCA 的 8 个区县,和新都区、东区共 10 个	广安市等东北部数个区县、眉山市和等成都周边区县、宜宾县等东南区县和西昌市共 34 个	东北部大部分区县、成都周围第二等级区县的外围区县、兴文县等东南区县,及川南,共 68 个	阿坝州等广大川西和川北高原地区,以及西昌周围的多数南部区县,共 69 个
NLM-SVM (C=30, γ=0.5)	同上	区县的地理分布同上,共 33 个	基本分布同上,但东南区县有差异,共 53 个	分布同上,共 85 个
SOFM	上述 10 个区县,及温江区、龙泉驿区、郫县、旌阳区、绵竹市,共 15 个	广安市等东北部大部分区县、眉山市和都江堰市等成都周边区县、宜宾县等东南区县和西昌市,共 42 个	万源市等少数东北部区县、峨眉山市等中南部、兴文县等东南区县,及川南,共 67 个	分布同上,共 57 个

(2) 基于 NLM 和 NLM—SVM 的分类结果能准确揭示出成都及其周边的经济发达区,以及川东南、攀西地区和川东北的经济密集区,其区别主要体现在川东北和川东南少数区县的分类上,但都能体现各区县的经济发展水平和整体的空间布局;同时 NLM 降维后各区县都有一维坐标,各分类等级内部目标可以通过该坐标来比较发展水平。

(3) 与 NLM 方法相比, SOFM 的差异主要表现在川东北地区的第二和第三等级的划分、中南部的第三和第四等级的划分、东南地区的第二和第三等级的划分上,而第一等级的个数也存在一些差异。SOFM 将成都北部德阳市的旌阳区和绵竹市,以及成都的温江区、龙泉驿区、郫县这 5 个区县被划分至第一等级;在第二等级上, SOFM 的分类结果添加了东北地区的一部分区县;其第三等级的结果与 NLM 的数目相同,但东北部分布有所差异,即东北部的目标大部分被划至第二类,而增加了中南地区的几个区县;第四等级的数目比 NLM 少 12 个,但分布大致相同,主要差异体现在中南地区。SOFM 的结果能符合四川经济发展的实际情况,并且与 NLM 的结果非常类似;但在东北部,由于遂宁—南充—广安地区形成了一个经济集中区,其经济相比周边具有一定的优势,这种格局在 SOFM 的分类结果中未得到体现,所以从这点考虑, SOFM 的结果有一定的错分情况;并且由于其结果来自于目标能否映射到同一个输出结点,是一个绝对的分类号,所以类内目标之间不能进行比较,这也是该方法的不足之处。

(4) 采集于 PCA 和 NLM 的两个不同已知样本集的 SVM 分类结果之间的差异比较

大。在 PCA-SVM 中，第二等级区县的数目非常少，而东北多数区县被划分至第四等级，并与 PCA 的分类结果相似，都未能准确展现四川经济发展现状；而 NLM-SVM 的结果与 NLM 类似，都能体现出经济发展的核心区域。由这两者结果之间的较大差异可知，已知小样本集的改变对结果有较大的影响，所以需要进一步考虑样本集的合理选择；最优参数的获取是一个区间搜索过程，不仅较难获取最优参数，同时效率也比较低。

从四种方法的分类结果分析可得知，这些结果之间的差异其实是一种“位移”现象，即揭示的总趋势是一致的，但是分界线的位置存在位移，这导致不同结果的差异较大。所有结果在体现整体趋势上都具有一定的正确性，但各算法原理的不同，导致结果之间存在较大差异，通过与实际情况的比对和分析，NLM 算法最能准确揭示四川经济发展水平及其空间格局，也能进行类内间的比较；SOFM 存在一定的错分情况；PCA 的结果与实际情况差异较大，但分值能体现总体发展情况；而 SVM 在没有已知样本集时的采样具有主观性，并且对结果影响较大，同时最优参数的搜索是一个复杂的过程。表 2 分析和总结了各种方法的优缺点，这为以后的应用和改进奠定了基础。

表 2 降维算法的比较分析

Tah 2 The comparison of multidimensional-reduction algorithms

算法	优点	缺点
PCA	理论完善、概念简单、计算方便、具有最优线性重构误差 ^[32] ；分值能揭示目标的分类总体趋势，能对类内目标进行比较	对非常高维的数据特征向量的计算可能不可行，主成分个数的确定没有明确的准则，不能用于处理非线性数据 ^[32] ；分类结果未能准确反映实际情况
NLM	能较好保持数据之间的相似性和差异性 ^[8] ；分类结果符合实际情况；降维坐标使得类内目标之间能进行比较	对非常高维的大数据集的计算代价大 ^[5] ；采用欧氏距离，在一定程度上不能揭示数据内在结构
SOFM	能保持数据的拓扑关系 ^[11] ；结果比较符合实际情况；	初始权值矢量、算法的学习率和获胜邻域对学习效果有影响；各输入模式类别可能有一定程度的混淆，分类精度不高 ^[33] ，结果存在一定的错误；难以进行类内目标的比较
SVM	坚实的数学理论基础、能避免“维数灾难”、具有很好的泛化性能、算法效率高 ^[9] ；需要较少的已知样本即可	参数众多，需反复调节参数才能得到最优结果；监督分类，需导入已知样本集对数据集进行训练，这使得该方法高度依赖于其他方法的结果，并且在选择样本时具有主观性

5 结论

考虑到不同降维算法由于其数学理论依据和适用范围的不同，最终得到的结果也必定具有差异，并且线性方法在分析非线性多维数据集时具有局限性，作者选择并实现了 PCA、NLM、SOFM 和 SVM 四种方法，并以四川 2007 年经济统计数据为例；然后基于四川经济发展的实际情况，对各分类结果进行分析，以说明各方法的优劣。PCA 算法求取的各区县分值虽然能在整体上揭示四川发展水平，但分类结果不能准确识别成都周边和川南环线的经济发展核心区域，在川东北区域也存在错分情况；NLM 能很好地体现出四川五大经济区域，在各区域中也能揭示出核心发展区域，能准确地体现四川经济发展的空间格局；SOFM 也能较好说明发展现状，但主要东北部存在一定的错分情况，并且难以

进行类内目标的比较;而 SVM 是监督分类,需要已知样本来训练分类过程,在样本的选择上存在很大的主观性,并且最优参数的搜索过程较为复杂。

总之,各种方法都具有一定的优势,同时也有不足之处。SVM 具有很好的数学理论基础,也能够最大化各类之间的距离,对具有已知样本集的数据能很好进行分析,但在未知样本中选择已知小样本集时具有一定的主观性,需要研究怎样更合理地选择样本集;NLM 方法中采取的是欧氏距离,这在一定程度上忽略了高维数据的潜在流形结构,而 ISOFMAP 等流形学习方法能克服这种不足,这也是下一步要进行的工作。

参考文献:

- [1] 吴晓婷,闫德勤. 数据降维方法分析与研究. 计算机应用研究,2009,26(8):2832~2835.
- [2] 丛明珠,欧向军,赵清,等. 基于主成分分析法的江苏省土地利用综合分区研究. 地理研究,2008,27(3):574~582.
- [3] 张吉献. 基于主成分分析法的河南省各城市综合实力评价. 河南科学,2009,27(1):115~118.
- [4] Borg I, Goren P. Modern Multidimensional Scaling: Theory and Application. Springer: New York, 1997.
- [5] Agrafiotis D K, Rassokhin D N, Lobanov V S. Multidimensional scaling and visualization of large molecular similarity tables. Journal of Computational Chemistry, 2001, 22(5): 488~500.
- [6] Schölkopf B, Smola A, Müller K R. Nonlinear component analysis as a kernel eigenvalue problem. Neural Computation, 1998, 10(5): 1299~1319.
- [7] Sammon J W. A nonlinear mapping for data structure analysis. IEEE Transactions on Computers, 1969, 18(5): 401~409.
- [8] 邵超. 非线性降维技术的研究及其在数据可视化中的应用. 北京: 北京交通大学博士学位论文, 2006.
- [9] 高隽. 神经网络原理及仿真实例. 北京: 机械工业出版社, 2003.
- [10] 阎平凡, 张长水. 神经网络与模拟进化计算(第二版). 北京: 清华大学出版社, 2005.
- [11] Teuvo Kohonen. Self-organized formation of topologically correct feature maps. Biological Cybernetics, 1982, 43(1): 59~69.
- [12] Juha Vesanto. SOM-Based data visualization methods. Intelligent Data Analysis, 1999, 3(2): 111~126.
- [13] 江波, 张黎. 基于多维自组织特征映射的聚类算法研究. 计算机科学, 2008, 35(6): 181~185.
- [14] Cortes C, Vapnik V. Support-vector networks. Machine Learning, 1995, 20: 273~297.
- [15] Christopher J C Burges. A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery, 1998, 2(2): 121~167.
- [16] Steve Gunn. Support Vector Machine for Classification and Regression. ISIS Technical Report, 1998.
- [17] 张学工. 关于统计学习理论与支持向量机. 自动化学报, 2000, 26(1): 32~42.
- [18] 罗公亮. 核函数方法(上). 冶金自动化, 2002, (3): 1~4.
- [19] 罗公亮. 核函数方法(下). 冶金自动化, 2002, (4): 1~4.
- [20] 王源, 陈亚军. 基于核的支持向量机构造方法的研究. 微机发展, 2005, 15(12): 96~98.
- [21] Hsu Chih-Wei, Chang Chih-Chung, Lin Chih-Jen. A Practical Guide to Support Vector Classification. <http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf>, 2010-08-03.
- [22] 王睿. 关于支持向量机参数选择方法分析. 重庆市大学学报: 自然科学版, 2007, 24(2): 36~39.
- [23] 王鹏, 朱小燕. 基于 RBF 核的 SVM 的模型选择及其应用. 计算机工程与应用, 2003, 24: 72~73.
- [24] Chapelle O, Vapnik V, Busquet O, et al. Choosing multiple parameters for support vector machines. Machine Learning, 2002, 46(1-3): 131~159.
- [25] 李琳, 张晓龙. 基于 RBF 核的 SVM 学习算法的优化计算. 计算机工程与应用, 2006, 29: 190~192.
- [26] 董国君, 哈力木拉提. 买买提, 等. 基于 RBF 核的 SVM 核参数优化算法. 新疆大学学报: 自然科学版, 2009, 26(3): 355~358.
- [27] 王如渊, 李翠华, 张学辉, 等. 四川省 FDI 区位选择的特征与机理. 地理研究, 2008, 27(2): 385~396.
- [28] 陈钊. 四川重点区域发展战略研究. 西华大学学报: 哲学社会科学版, 2005, 4(3): 17~20.
- [29] 李斌, 董锁成, 李雪. 四川省生态经济区划研究. 四川农业大学学报, 2009, 27(3): 302~308.

- [30] 张杰. 川渝经济发展水平的比较研究. 重庆工学院学报, 2006, 20(7): 47~49.
- [31] 张杰. 重庆、四川主要经济指标的比较研究. 重庆工商大学学报: 西部论坛, 2006, 16(3): 43~45.
- [32] 吴晓婷, 闫德勤. 数据降维方法分析与研究. 计算机应用研究, 2009, 26(8): 2832~2835.
- [33] 梁斌梅. 自组织特征映射神经网络的改进及应用研究. 计算机工程与应用, 2009, 45(31): 134~137.

Study on dimension-reduction of spatial economic statistics: A case study of economic statistical data of Sichuan

DONG Cheng-wei^{1, 2}, RUI Xiao-ping^{1, 3}, DENG Yu^{4, 5, 6}, GUAN Xing-liang^{4, 5}

(1. College of Resources and Environment, Graduate University of Chinese Academy of Sciences, Beijing 100049, China; 2. Beijing Institute of Surveying and Mapping, Beijing 100038, China; 3. Research Center for Eco-Environmental Sciences, CAS, Beijing 100085, China; 4. Institute of Geographic Sciences and Natural Resources Research, CAS, Beijing 100101, China; 5. Graduate university of Chinese Academy of Sciences, Beijing 100049, China; 6. Harvard University, Cambridge 02138, USA)

Abstract: There are more than three attributes in economic statistical data generally. When studying the inherent structural characteristics of these data such as clustering and distribution, researchers need to reduce multi-dimensional information to three-dimensional space or less to achieve multi-dimensional visualization. There are multi-dimensional reduction methods, whose results are different from each other because of different mathematics theories and application ranges, and the visualization results of these methods will vary. So evaluation of different methods can provide important references for the selection of methods in different areas. In the paper, the authors analyze economic statistical data of Sichuan province in 2007 based on county-unit by implementing four commonly used algorithms: the linear method PCA, nonlinear method NLM and SOFM, and a supervised classification method SVM, then obtain a series of classification results. Considering the status of economic development in Sichuan, the authors analyze the differences between the results of these methods, and draw some conclusions as follows. Although PCA can reveal the overall development trend, the result is not consistent with the real condition in Sichuan; NLM can well show the regional trend and core areas of economic development in Sichuan, and account for the development status; SOFM can also show the development status, but there are several classification errors in the northeastern part of the region. It is impossible for comparison within each cluster; as a supervised method, SVM needs a known sample set to train the classification process, which makes the sample selection subjective, and the search process for optimal parameters is complicated. The comparison of these methods and their application in economic statistics fields can provide a reference for the future relevant spatial dimension-reduction research.

Key words: dimension-reduction; multi-dimensional visualization; economic statistics data; Sichuan