HZ BOOKS 华章IT

[PACKT] PUBLISHING

数据科学与工程技术丛书

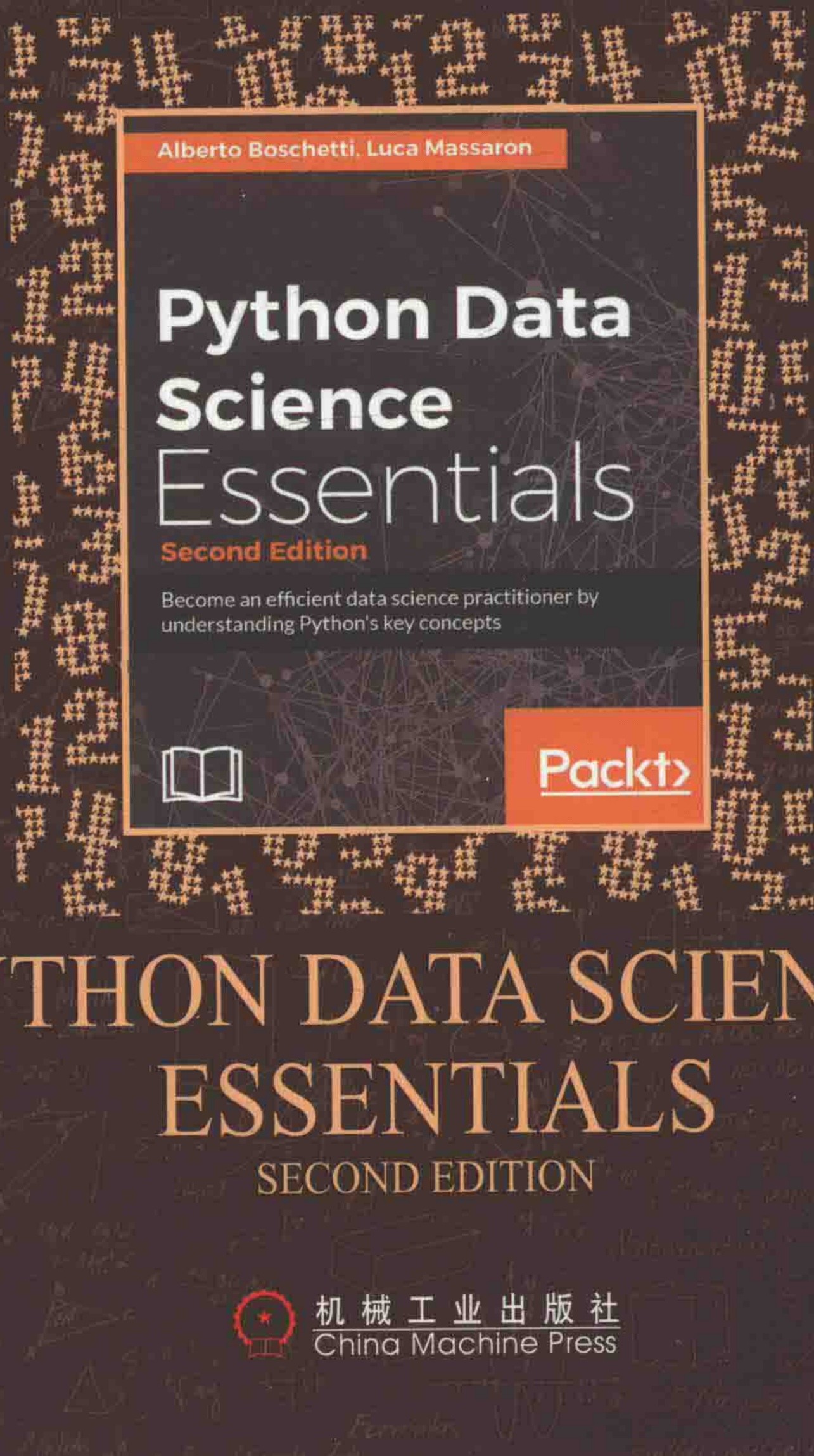
数据科学导论

Python语言实现

（原书第2版）

[意] 阿尔贝托·博斯凯蒂（Alberto Boschetti）
卢卡·马萨罗（Luca Massaron） 著

于俊伟 靳小波 译



PYTHON DATA SCIENCE ESSENTIALS

SECOND EDITION

机械工业出版社
China Machine Press

数据科学与工程技术丛书

PYTHON DATA SCIENCE ESSENTIALS
SECOND EDITION

数据科学导论

Python语言实现

（原书第2版）

[意] 阿尔贝托·博斯凯蒂（Alberto Boschetti）
卢卡·马萨罗（Luca Massaron） 著

于俊伟 靳小波 译

机械工业出版社
China Machine Press

图书在版编目（CIP）数据

数据科学导论：Python 语言实现（原书第 2 版）/（意）阿尔贝托·博斯凯蒂（Alberto Boschetti），（意）卢卡·马萨罗（Luca Massaron）著；于俊伟，靳小波译．—北京：机械工业出版社，2018.1
（数据科学与工程技术丛书）
书名原文：Python Data Science Essentials，Second Edition

ISBN 978-7-111-58986-0

I. 数… II. ①阿… ②卢… ③于… ④靳… III. 软件工具－程序设计 IV. TP311.561

中国版本图书馆 CIP 数据核字（2018）第 011393 号

本书版权登记号：图字 01-2017-2751

本书首先介绍了如何在 Python 3.5 中安装必要的数据科学工具箱；然后引导你进入数据改写和预处理阶段，在其中阐述用于数据分析、探索或处理的数据加载、变换和修复等关键的数据科学活动；最后，通过介绍主要的机器学习算法、图分析技术和可视化方法来对数据科学进行概述。

出版发行：机械工业出版社（北京市西城区百万庄大街 22 号　邮政编码：100037）

责任编辑：张志铭	责任校对：李秋荣
印　　刷：中国电影出版社印刷厂	版　　次：2018 年 3 月第 1 版第 1 次印刷
开　　本：185mm×260mm　1/16	印　　张：15（含彩插 0.5 印张）
书　　号：ISBN 978-7-111-58986-0	定　　价：59.00 元

凡购本书，如有缺页、倒页、脱页，由本社发行部调换

客服热线：（010）88378991　88361066　　投稿热线：（010）88379604

购书热线：（010）68326294　88379649　68995259　　读者信箱：hzjsj@hzbook.com

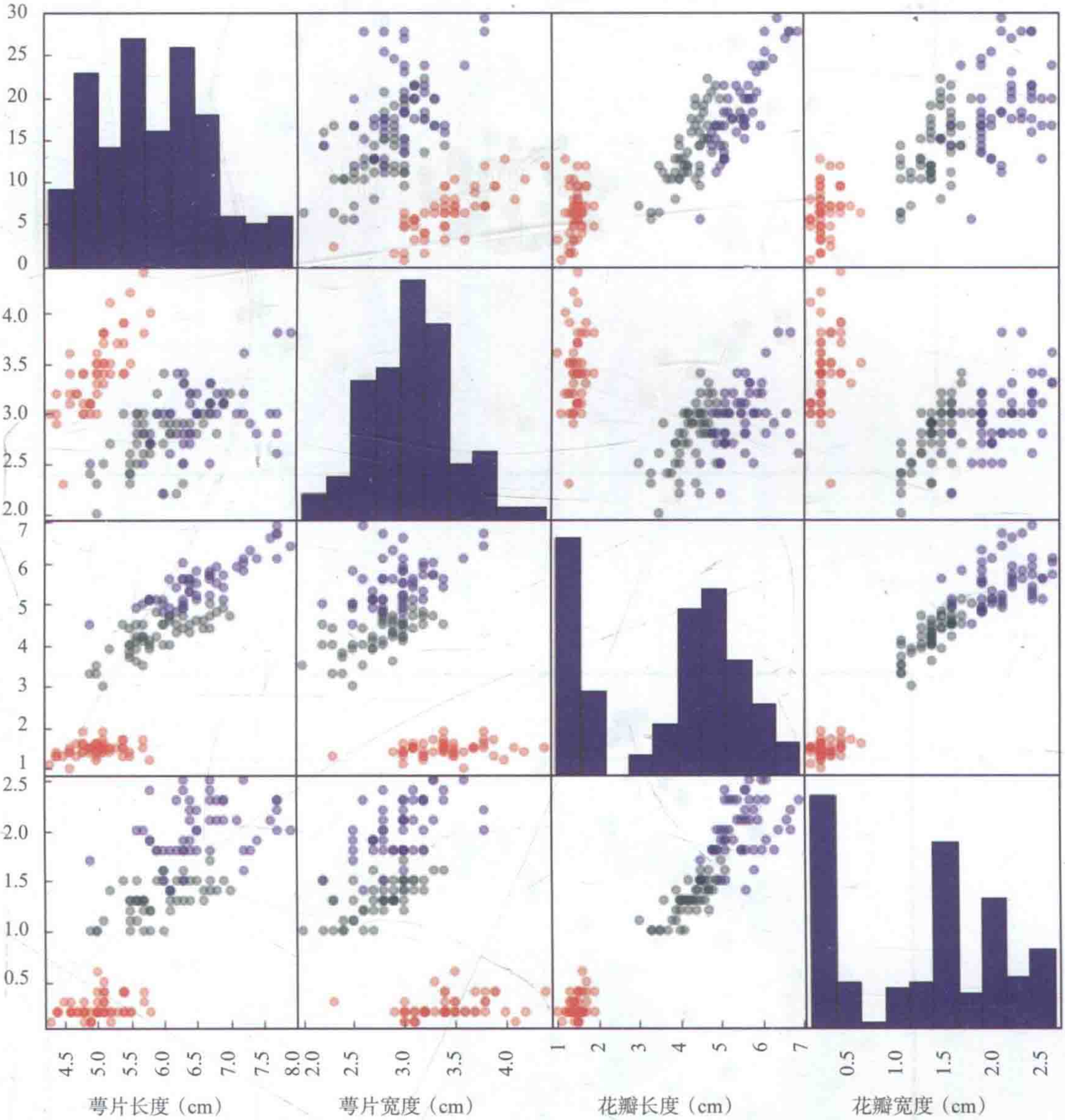


图 1

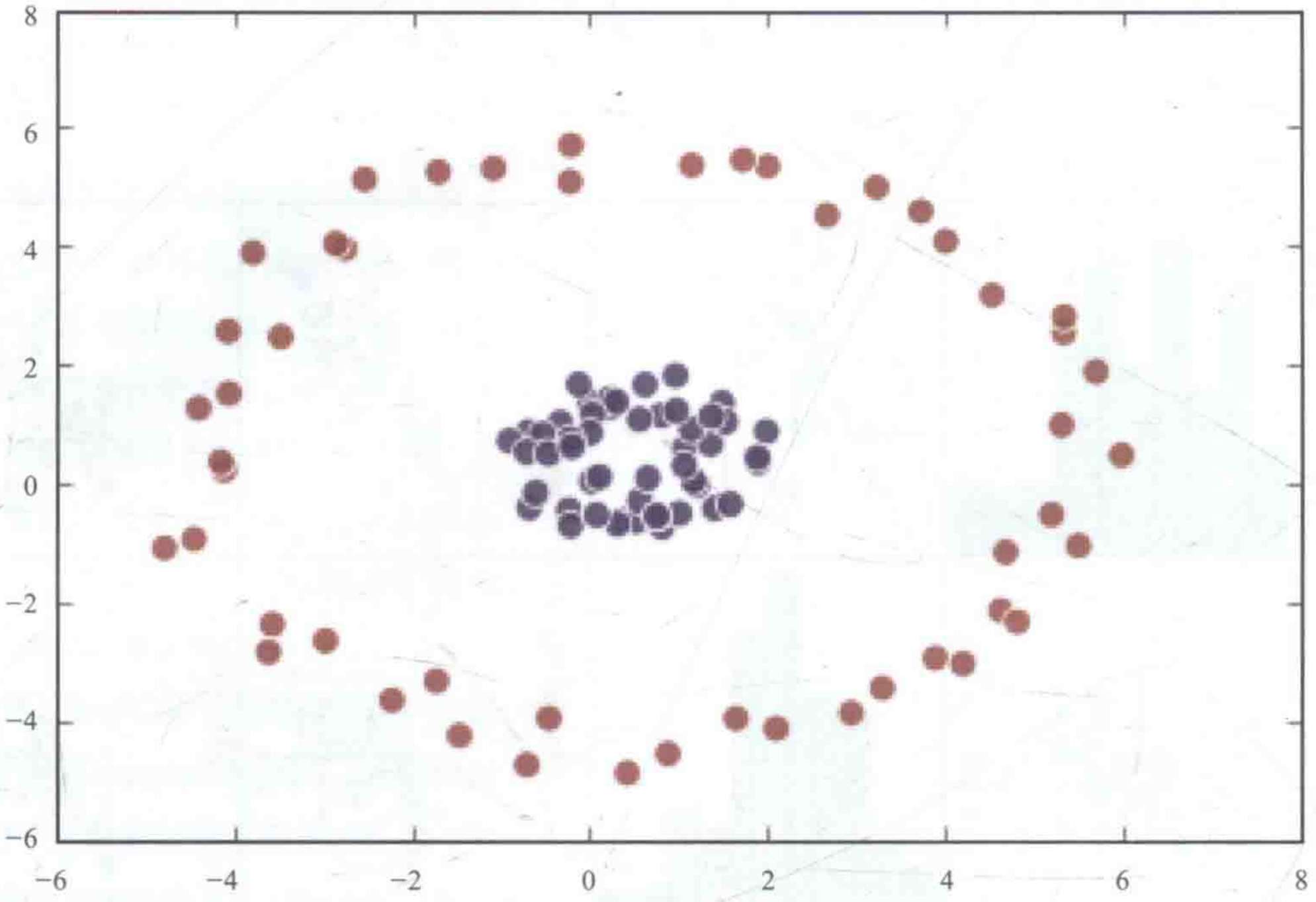


图 2

图 3

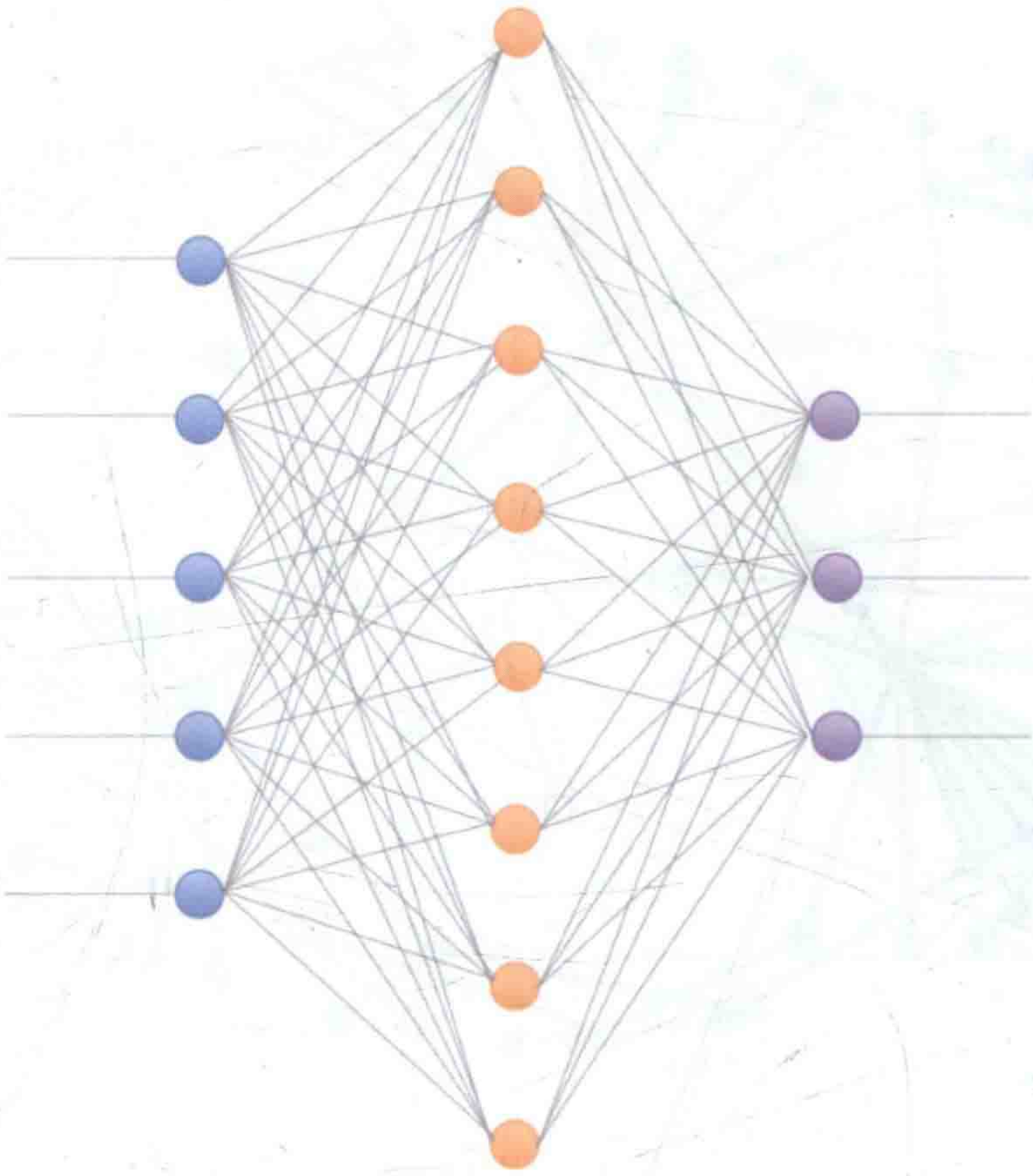


图 4

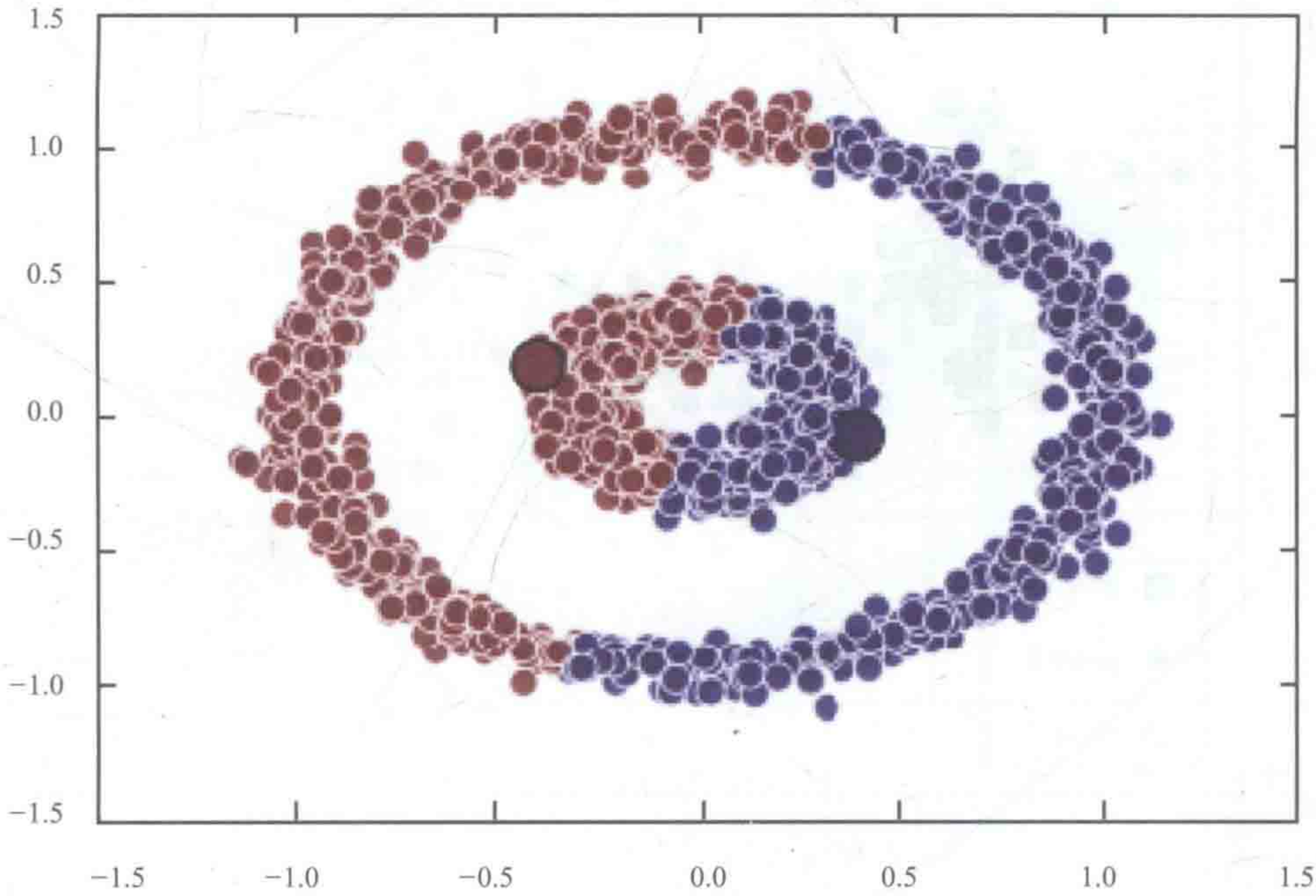


图 5

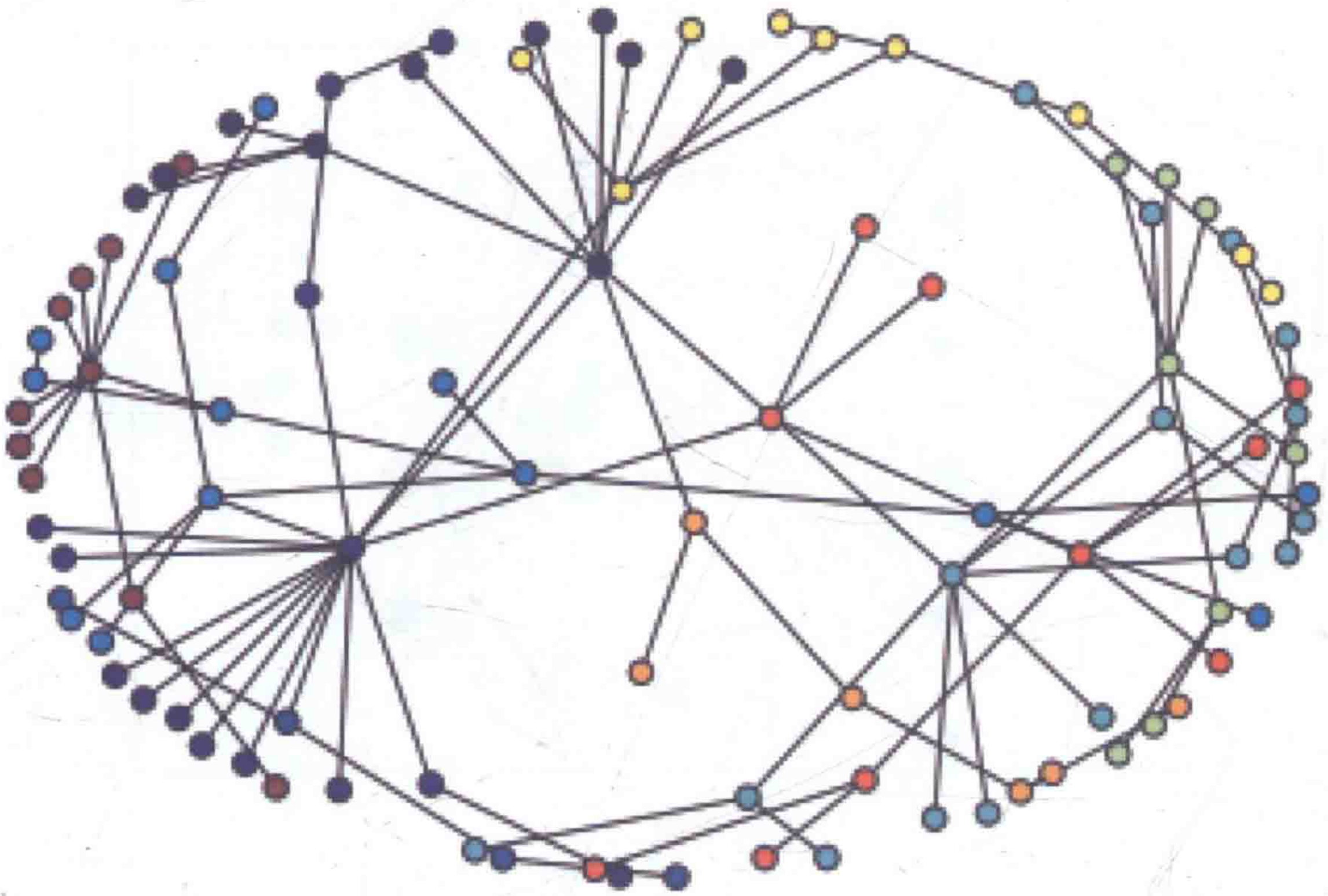


图 6

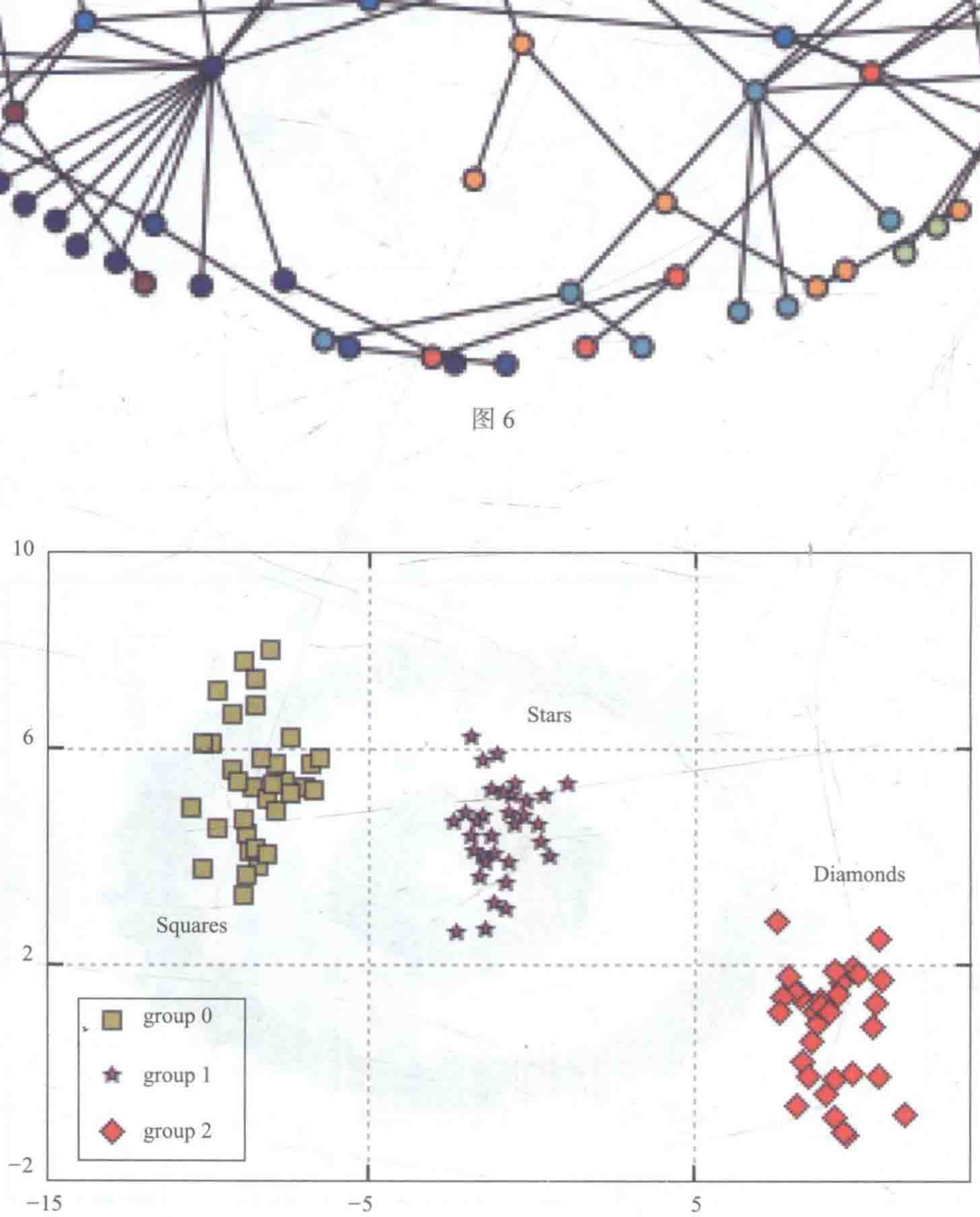


图 7

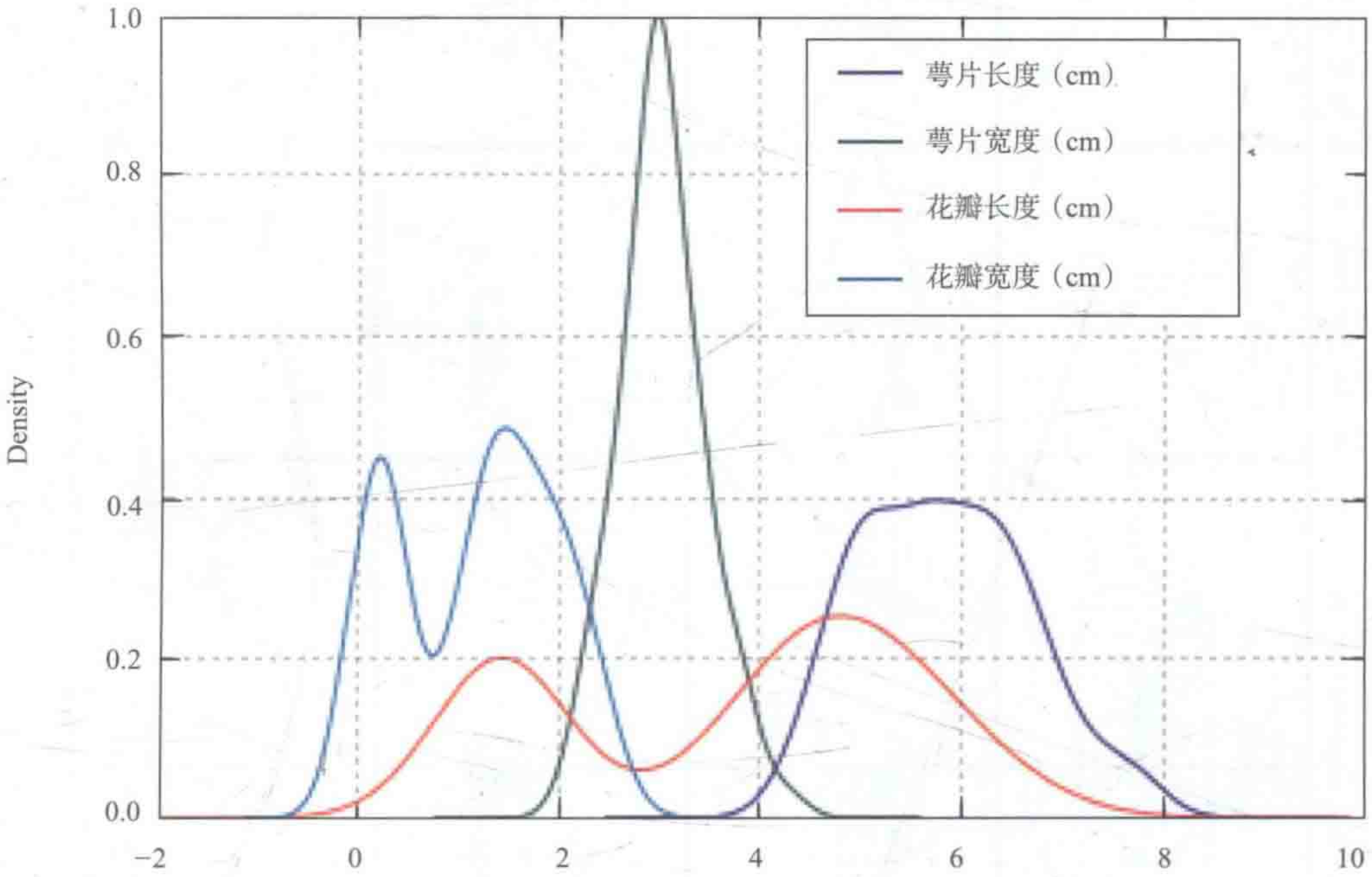

1.0
0.8
0.6
0.4
0.2
0.0
Density
−2
0
2
4
6
8
10
萼片长度（cm）
萼片宽度（cm）
花瓣长度（cm）
花瓣宽度（cm）

图 8

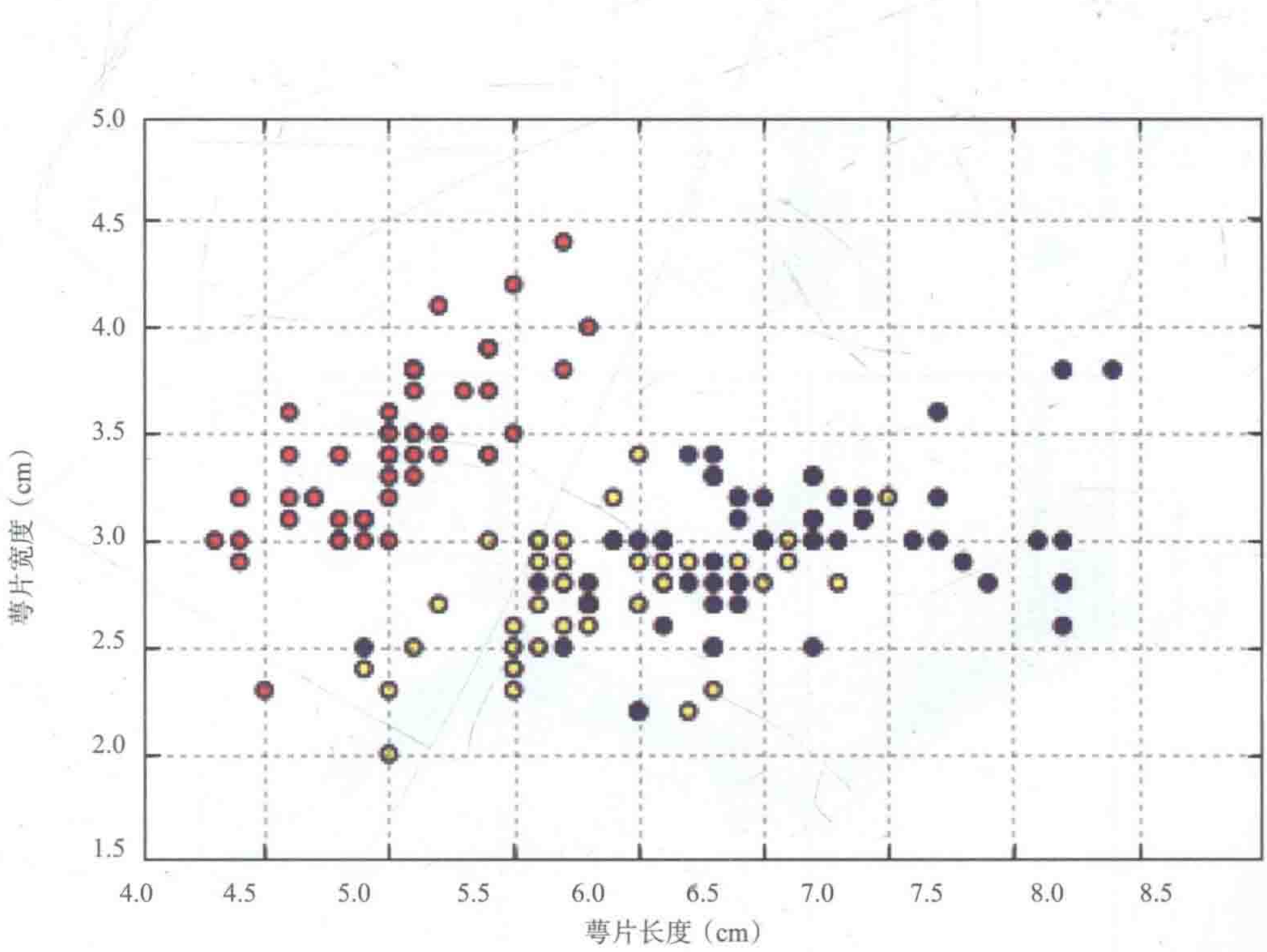

5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
萼片宽度（cm）
4.0
4.5
5.0
5.5
6.0
6.5
7.0
7.5
8.0
8.5
萼片长度（cm）

图 9

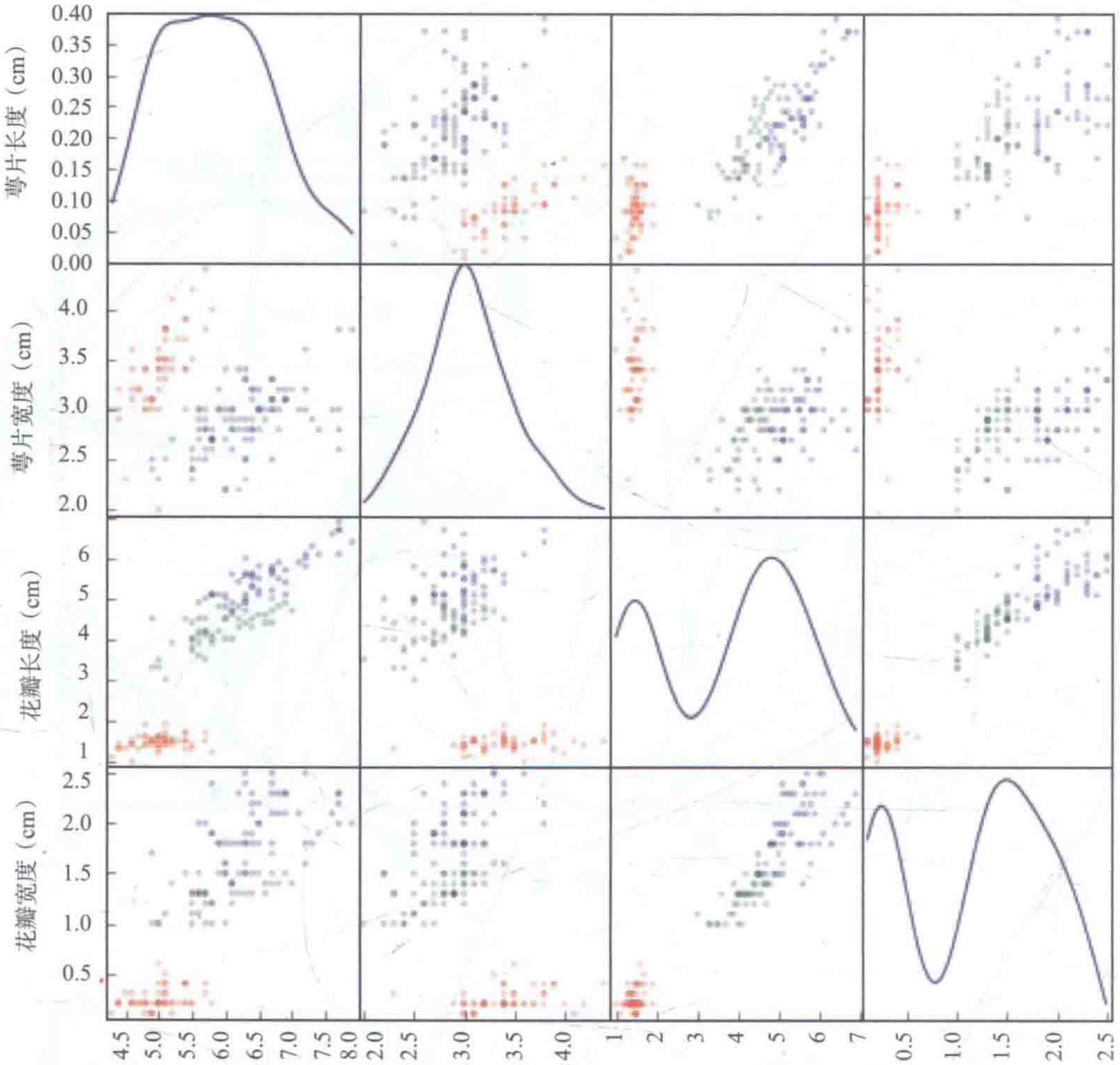

萼片长度（cm）
萼片宽度（cm）
花瓣长度（cm）
花瓣宽度（cm）
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
4.0
3.5
3.0
2.5
2.0
6
5
4
3
2
1
2.5
2.0
1.5
1.0
0.5
4.5
5.0
5.5
6.0
6.5
7.0
7.5
8.0
2.0
2.5
3.0
3.5
4.0
1
2
3
4
5
6
7
0.5
1.0
1.5
2.0
2.5
萼片长度（cm）
萼片宽度（cm）
花瓣长度（cm）
花瓣宽度（cm）

图 10

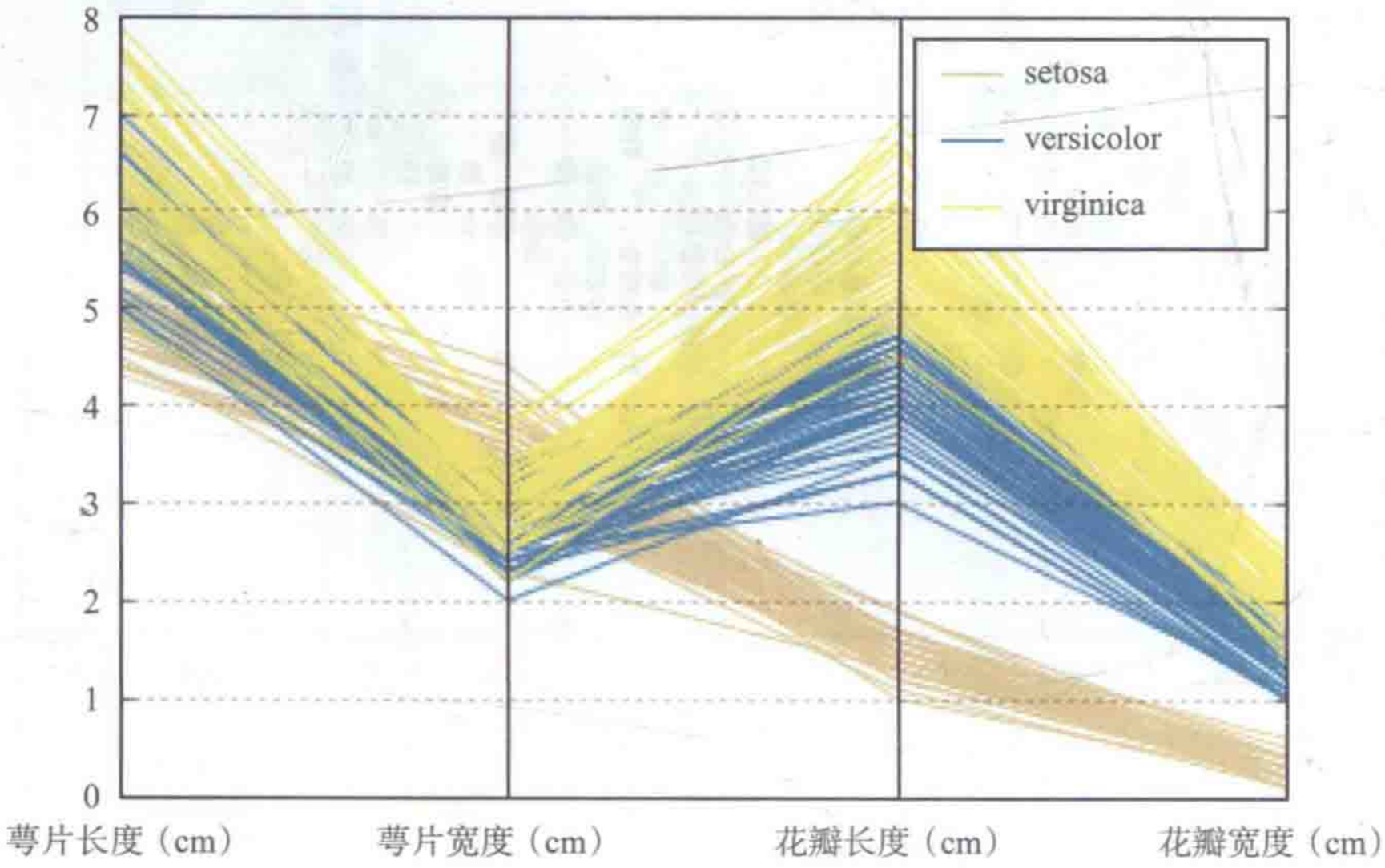

8
7
6
5
4
3
2
1
0
setosa
versicolor
virginica
萼片长度（cm）
萼片宽度（cm）
花瓣长度（cm）
花瓣宽度（cm）

图 11

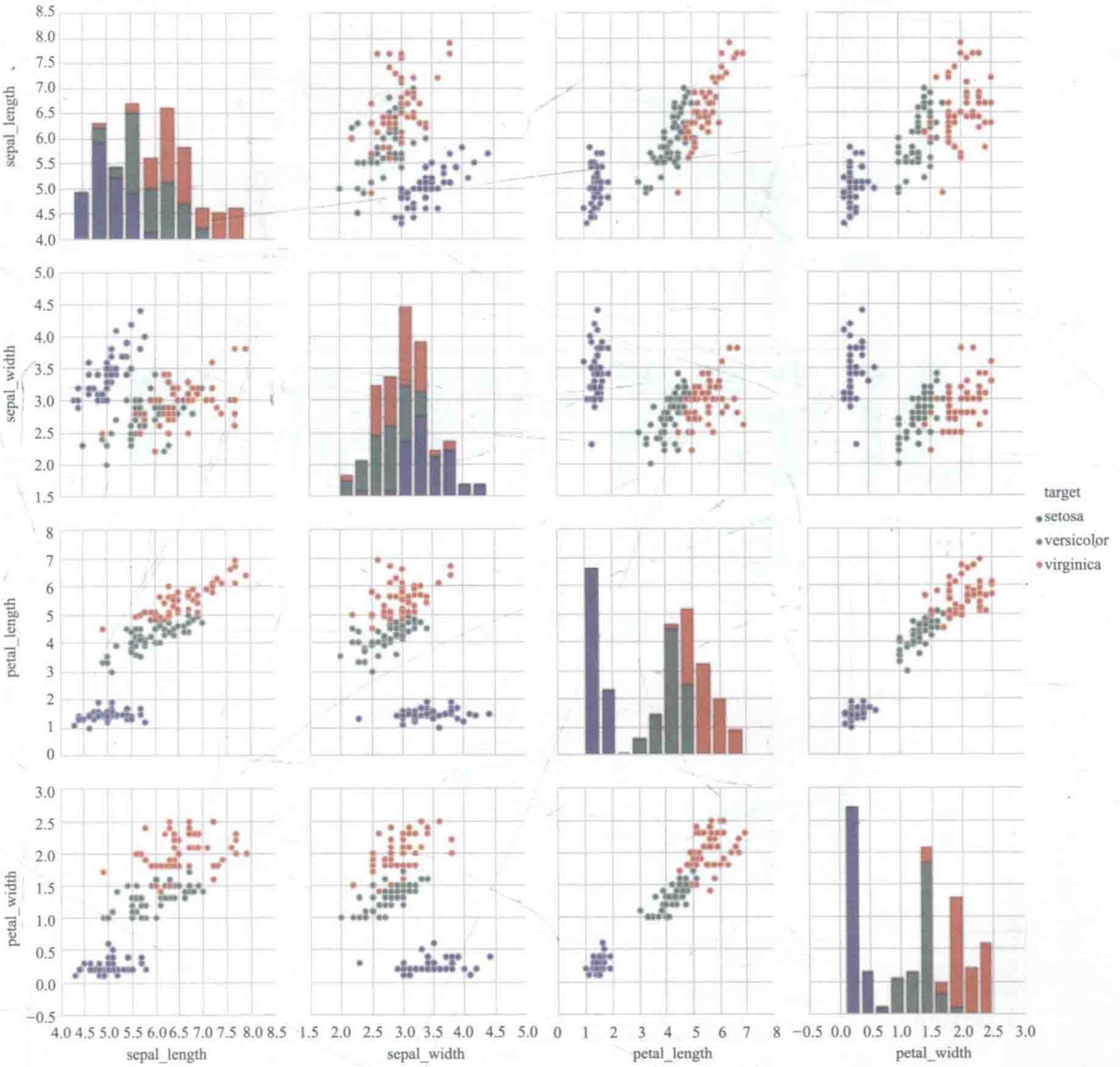

sepal_length
sepal_width
petal_length
petal_width
target
setosa
versicolor
virginica

图 12

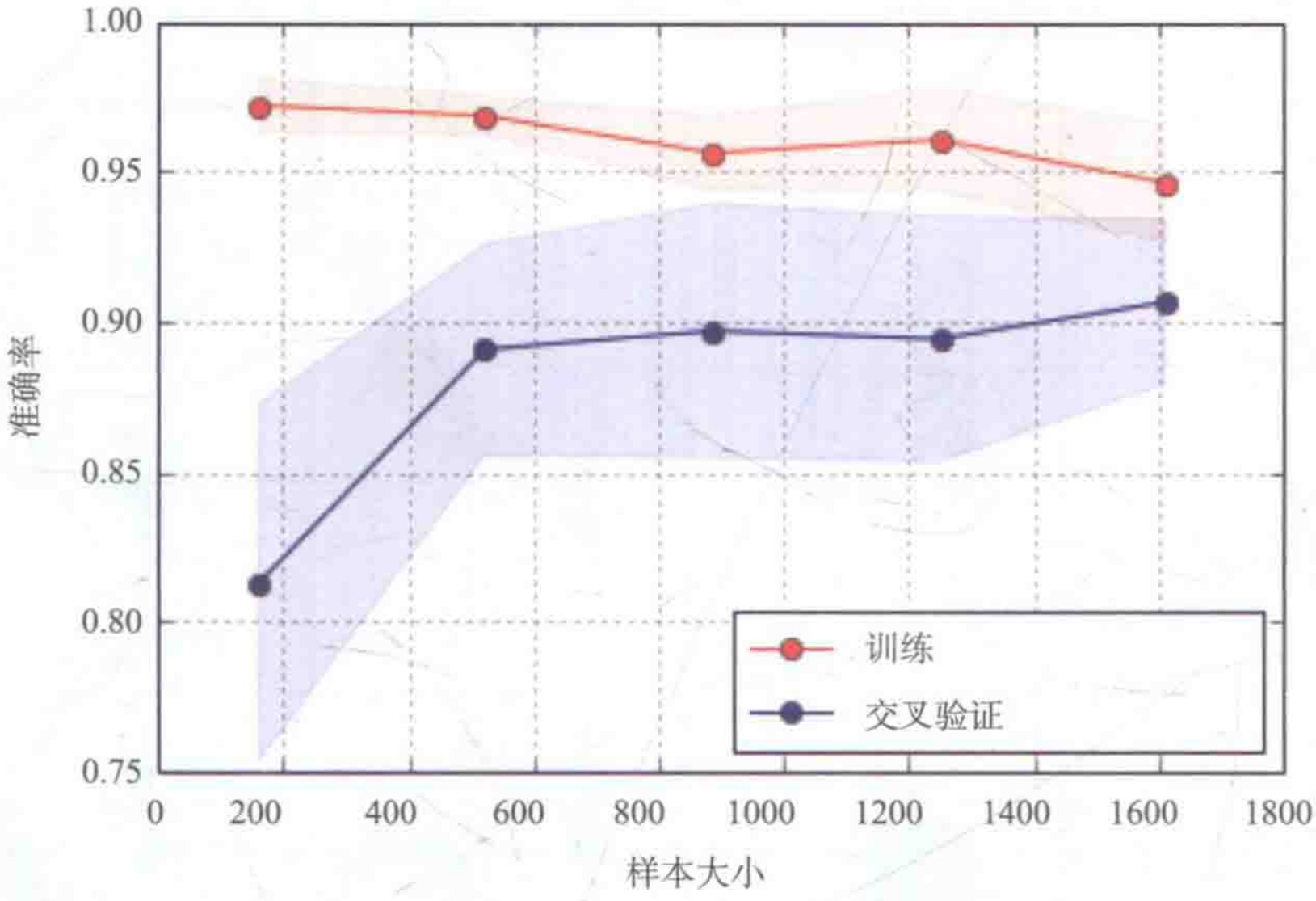


图 13

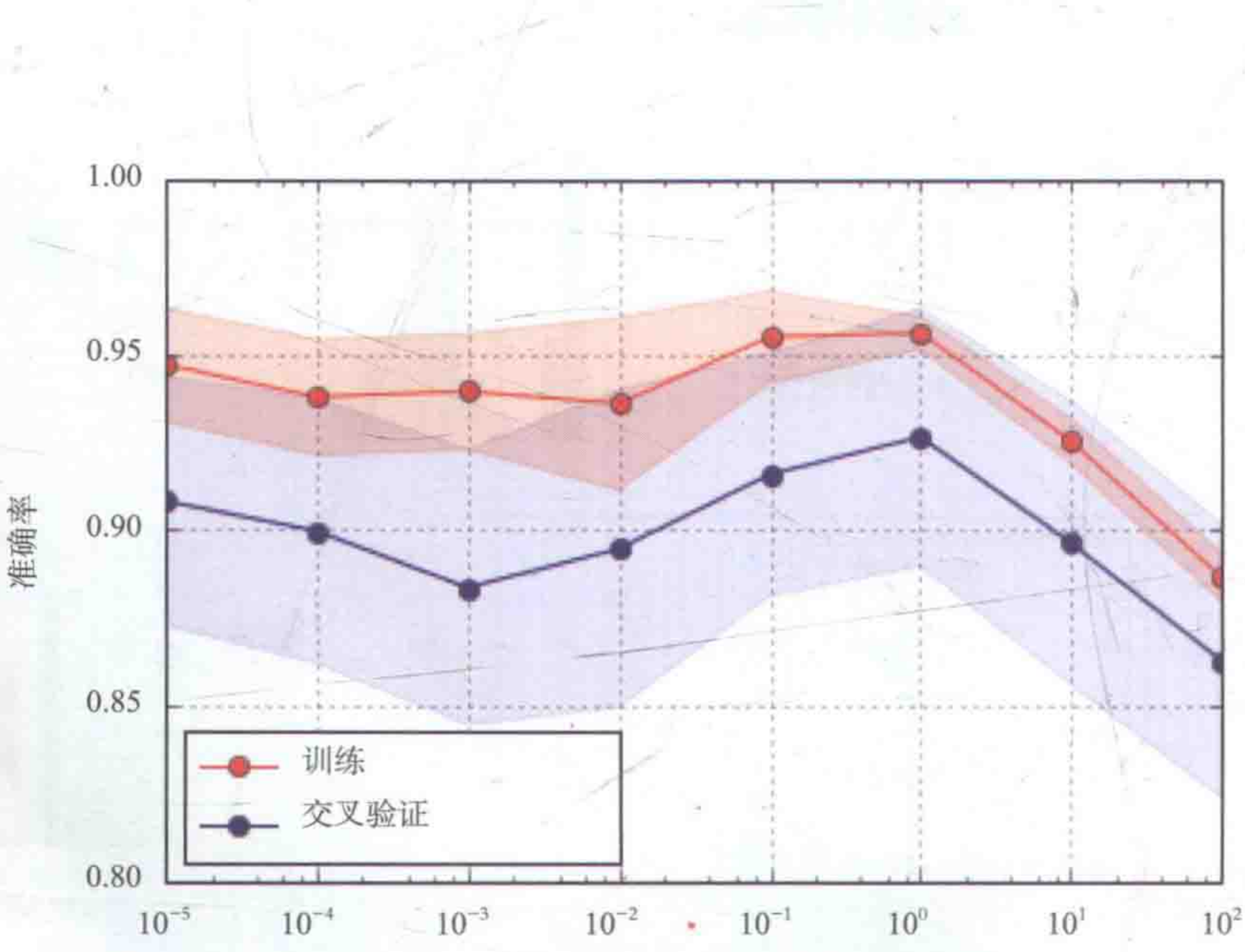


图 14

译 者 序

我们正处于一个快速发展的信息化时代，人们每天都在生产着各种类型的数据，与此同时，数据也在极大地影响着我们的生活。于是，数据成为与能源同等重要的资源。掌握了数据获取、处理、建模、分析等过程的理论和方法，无疑就掌握了打开这种新型资源的钥匙。

数据科学是融合多种学科的新的知识领域，一般要求学习者或从业者具备统计学等数学知识、计算机相关学科专业知识和特定业务领域的知识。随着人工智能、大数据、云计算等领域的蓬勃发展，众多科技公司和从业人员都越来越重视对数据科学的研究和应用。

工欲善其事，必先利其器。那么，什么才是数据科学家最值得信赖的专业工具呢？Python 无疑是众多数据分析语言中最适合的一个。GitHub Octoverse 2017 年度报告显示，Python 已经代替 Java 成为第二种最受欢迎的编程语言，由此可见数据科学发展的火热程度。Python 是一种通用的、解释性和面向对象的语言，具有强大的数据分析和机器学习软件包，为解决各种数据科学问题提供了快速、可靠、成熟的开发环境。它易学易用，便于快速开发，有很好的交互式体验，并且已经征服了科学界，堪称解决数据科学问题的神器。

本书介绍了进行数据科学分析和开发的所有关键要点，包括 Python 软件及相关工具包的安装和使用；介绍了数据加载、运算和改写等基本数据准备过程，以及特征选择、维数约简等高级数据操作方法；建立了由训练、验证、测试等过程组成的数据科学流程，结合示例深入浅出地讲解了多种机器学习算法；讨论了基于图模型的社交网络创建、分析和处理方法；最后还谈到了数据分析结果的可视化及相关工具使用方法。

本书在第 1 版的基础上对数据科学的编程环境和知识结构进行了一定的更新和调整，方便读者使用最新版本的 Python 及相关工具包，也有助于读者结合数据科学的最新发展形成自己的知识体系。在编程环境方面，本书的示例代码全部是基于 Python 3 的，尽管通过添加导入过程还能在 Python 2 环境下运行，我们还是建议读者使用代表未来发展趋势的 Python 3 的较新版本。Jupyter 已经正式从 IPython 分离，本书对这种支持多语言的交互式工具进行了系统而详细的介绍，书中很多示例都是在 Jupyter Notebook 环境下完成的。另外，为了给不同系统和不同版本的 Python 创建独立的运行环境，可以使用

virtualenv 或 conda 进行虚拟环境管理。在知识结构上，本书对数据处理、机器学习和结果表示等相关的方法进行了梳理和补充。例如，将目前流行的大数据和深度学习单独列为一节进行详细介绍，重点介绍了深度学习库 Theano 和神经网络 Keras 库的使用方法。为了促进机器学习方法在现实生活中的应用，6.4 节介绍了创建“机器学习即服务”预测服务器的工具和方法。另外，为了增强读者的 Python 基础，附录简明扼要地介绍了 Python 数据结构、函数、类、条件语句等基础知识，还指出了深入学习 Python 相关知识会用到的网络资源。

本书作者是两位意大利数据科学专家，他们长期从事与数据科学相关的教学和科研工作，在 Python 社区、社交网络上也很活跃，发表了多篇学术论文，出版了多部著作，对数据科学相关人员影响很大。本书是作者多年实践经验的总结，具有以下特点：1）循序渐进，深入浅出，让初学者不畏惧，让从业者得要领。2）理论与实践相结合，几乎所有算法和理论都辅以简洁的实例和说明，通过简单的几行代码即可验证。3）有助于深入理解数据科学概念，轻松进行理论扩展，快速建立自己的工程，从而做到学以致用，促进多种形式的科学研究和应用开发。

无论是作为数据科学和机器学习理论研究者的参考书，还是作为数据科学应用开发人员的工具书，抑或作为有志成为数据科学家的初学者的指导书，本书都能提供非常有价值的参考。本书还可以作为高等院校相关学科本科生或研究生的学习教材，特别适合从事数据科学、信息处理和机器学习等方向的研究生进行学习和参考。

本书第 4 章由河南工业大学信息科学与工程学院靳小波博士翻译，其余章节及附录由河南工业大学信息科学与工程学院于俊伟博士翻译。由于译者水平有限，加之时间仓促，错误和疏漏在所难免，恳请读者批评指正。

本书的翻译工作得到国家自然科学基金项目（61300123）、河南省高校重点科研项目（15A520012）和河南省研究生教育教学改革研究与实践项目（2017SJGLX046Y）的资助。感谢机械工业出版社华章公司对本书出版的高度重视，特别感谢和静、张志铭等编辑的讨论和帮助，他们的辛勤工作提高了本译著的质量。最后还要感谢父母、妻子及女儿等家人的爱和包容，他们的支持使我对目前从事的科技工作更加坚定和执着，也希望我们的工作能为读者在数据科学的成长道路上提供有益的帮助！

于俊伟

2017 年 10 月

前　言

“千里之行，始于足下。”

——老子（公元前 604—531[⊖]）

数据科学属于一门相对较新的知识领域，它成功融合了线性代数、统计建模、可视化、计算语言学、图形分析、机器学习、商业智能、数据存储和检索等众多学科。

Python 编程语言在过去十年已经征服了科学界，现在是数据科学实践者不可或缺的工具，也是每一个有抱负的数据科学家的必备工具。Python 为数据分析、机器学习和算法求解提供了快速、可靠、跨平台、成熟的开发环境。无论之前在数据科学应用中阻止你掌握 Python 的原因是什么，我们将通过简单的分步化解和示例导向的方法帮你解决，帮助你在演示数据集和实际数据集上使用最直接有效的 Python 工具。

作为第 2 版，本书对第 1 版内容进行了更新和扩展。以最新的 Jupyter Notebook（包括可互换内核，一个真正支持多种编程语言的数据科学系统）为基础，本书包含了 NumPy、pandas 和 Scikit-learn 等库的所有主要更新。此外，本书还提供了不少新内容，包括深度学习（基于 Theano 和 Tensorflow 的 Keras）、漂亮的数据可视化（Seaborn 和 ggplot）和 Web 部署（使用 bottle）等。本书首先使用单源方法，展示如何在最新版 Python（3.5）中安装基本的数据科学工具箱，这意味着本书中的代码可以在 Python 2.7 上重用。接着，将引导你进入完整的数据改写和预处理阶段，主要阐述用于数据分析、探索或处理的数据加载、变换、修复等关键数据科学活动。最后，本书将完成数据科学精要的概述，介绍主要的机器学习算法、图分析技术和可视化方法，其中，可视化工具将更易于向数据科学专家或商业用户展示数据处理结果。

本书内容

第 1 章介绍 Jupyter Notebook，演示怎样使用程序手册中的数据。

⊖ 目前国内比较认可的老子生卒年是公元前 571—471。译者有幸生于老子故里，对老子的传说和史料有所了解，但众多考证都只能给出一个大概的年限。这里译者对作者严谨的引述表示敬意，或许以后利用数据科学技术能从众多史料中挖掘出更确切的老子生平。——译者注

第 2 章对数据科学流程进行概述，详细分析进行数据准备和处理所使用的关键工具，这些工具将在采用机器学习算法和建立假设实验计划之前使用。

第 3 章讨论所有可能有助于结果改进甚至提升的数据操作技术。

第 4 章深入研究 Scikit-learn 包中的主要机器学习算法，例如线性模型、支持向量机、树集成和无监督聚类技术等。

第 5 章介绍图的概念，它可以表示为偏离预测或目标的有趣矩阵。这是目前数据科学界的研究热点，期待利用图的技术来研究复杂的社交网络。

第 6 章介绍使用 matplotlib 进行可视化的基本方法，以及如何使用 pandas 进行探索性数据分析（EDA），如何使用 Seaborn 和 Bokeh 实现漂亮的可视化，还包括如何建立提供所需要信息的 Web 服务器。

附录包括一些 Python 示例和说明，重点介绍 Python 语言的主要特点，这些都是从事数据科学工作必须了解的。

阅读准备

本书用到的 Python 及其他数据科学工具（从 IPython 到 Scikit-learn）都能在网上免费下载。要运行本书附带的源代码，需要一台装有 Windows、Linux 或 Mac OS 等操作系统的计算机。本书将分步介绍 Python 解释器的安装过程，以及运行示例所需要的工具和数据。

读者对象

如果你有志于成为数据科学家，并拥有一些数据分析和 Python 方面的基础知识，本书将助你在数据科学领域快速入门。对于有 R 语言或 Matlab 编程经验的数据分析人员，本书也可以作为一个全面的参考书，提高他们在数据操作和机器学习方面的技能。

代码下载

你可以从 http://www.packtpub.com 通过个人账号下载你所购买书籍的样例源码。你也可以访问华章图书官网 http://www.hzbook.com，通过注册并登录个人账号下载本书的源代码。

彩图下载

我们还提供了一个 PDF 文件，其中包含本书中使用的截图和彩图，可以帮助读者更好地了解输出的变化。文件可以从以下地址下载：http://www.packtpub.com/sites/default/files/downloads/PythonDataScienceEssentialsSecondEdition_colorImages.pdf。

作者简介

阿尔贝托·博斯凯蒂（Alberto Boschetti）

数据科学家、信号处理和统计学方面的专家。他是通信工程专业博士，现在伦敦居住和工作。他主要从事自然语言处理、行为分析、机器学习和分布式处理等方面的挑战性工作。他对工作充满激情，经常参加学术聚会、研讨会及其他学术活动，紧跟数据科学技术发展的前沿。

我要感谢我的家人、朋友和同事！同时，也非常感谢开源社区！

卢卡·马萨罗（Luca Massaron）

数据科学家、市场营销研究主导者，是多变量统计分析、机器学习和客户洞察方面的专家。有十年以上解决实际问题的经验，使用推理、统计、数据挖掘和算法为利益相关者创造了巨大的价值。在意大利他是网络受众分析的先锋，并在 Kaggler 上获得排名前十的佳绩，随后一直热心参与各种与数据及数据分析相关的活动，积极给新手和专业人员讲解数据驱动知识发现的潜力。他崇尚大道至简，坚信理解数据科学的精要能给你带来巨大收获。

致 Yukiko 和 Amelia，谢谢你们的爱和包容。“前路无止境，星云作伴长，双脚虽远行，终归还家乡。”

目　录