马克威统计分析与数据挖掘应用案例

MARKWAY

主编 黄 晖
李鸿琪

中国统计出版社
China Statistics Press

马克威统计分析与数据挖掘应用案例

MARKWAY

主编 黄 晖
李鸿琪

中国统计出版社
China Statistics Press

(京)新登字 041 号

图书在版编目(CIP)数据

马克威统计分析与数据挖掘应用案例/黄晖,李鸿琪主编. ——北京:中国统计出版社,2012.8
ISBN 978-7-5037-6620-6

Ⅰ.①马… Ⅱ.①黄… ②李… Ⅲ.①统计分析-统计程序 ②数据采集 Ⅳ.①C819 ②TP274

中国版本图书馆 CIP 数据核字(2012)第 173940 号

马克威统计分析与数据挖掘应用案例

作　　者/黄　晖　李鸿琪
责任编辑/梁　超
封面设计/孙　婷
出版发行/中国统计出版社
通信地址/北京市西城区月坛南街 57 号　邮政编码/100826
办公地址/北京市丰台区西三环南路甲 6 号　邮政编码/100073
网　　址/http://csp.stats.gov.cn
电　　话/邮购(010)63376907　书店(010)68783172
印　　刷/三河市利兴印刷有限公司
经　　销/新华书店
开　　本/710×1000mm　1/16
字　　数/360 千字
印　　张/20.25
印　　数/1—3000 册
版　　别/2012 年 9 月第 1 版
版　　次/2012 年 9 月第 1 次印刷
书　　号/ISBN 978-7-5037-6620-6/C.2697
定　　价/39.00 元

中国统计版图书,版权所有,侵权必究。
中国统计版图书,如有印装错误,本社发行部负责调换。

前　言

马克威分析系统是上海天律信息技术有限公司自主开发的中国第一套大型统计分析与数据挖掘软件。自诞生至今，马克威分析软件在政府、军队、企事业单位和社会团体得到了广泛的应用。为了使广大统计工作者更好地掌握统计分析和数据挖掘方法，熟悉马克威统计分析与数据挖掘软件的操作，在国家统计局的帮助下我们选编了“马克威统计分析与数据挖掘应用案例”。

本书包括十四篇应用案例，分别是“上海世博会客流量预测”、“扩展线性支出系统——上海市区居民消费结构实证分析”、“二值逻辑回归——马克威软件在调查问卷分析中的应用”、“误差修正模型——上海市居民消费实证分析”、“因子分析——长三角16城市社会经济发展研究”、“多元方差分析——笔记本电脑价格影响因素分析”、“面板数据模型——中国城镇居民消费研究”、“偏最小二乘回归——三次产业增加值与支出法GDP的关系分析”、“PLS路径模型——美国顾客满意度指数模型”、“RBF神经网络在上海市二手房价格评估中的应用”、“决策树在电信客户流失预测中的应用”、“贝叶斯网络在临床诊断乙肝中的应用”、“江城市宏观经济年度模型”、“北京市人口预测与分析”。

上述十四篇文章被安排成十四章。其中，第一章“上海世博会客流量预测”介绍了综合预测方法，最后一章“北京市人口预测与分析”介绍了人口预测与分析的一个案例，希望对全国第六次人口普查数据的分析有所帮助。其他各章都是通过一个案例介绍一种统计方法或数据挖掘方法，如居民消费结构实证分析介绍了扩展线性支出系统，毕业生调查案例介绍了二值逻辑回归，宏观经济年度模型介绍了联立方程、满意度调查介绍了PLS路径模型，等等。每一章的内容安排大体上分为三个部分。第一部分为统计方法或数据挖掘的原理介绍，该部分内容着重基本原理的介绍，不进行大量的公式推导。第二部分为案例分析，这

是每一章的重点，包括数据的组织、分析、结果的解读，等等。第三部分为马克威软件的操作，尽可能细致地教会学员如何方便地操作马克威分析系统。总的目的是：

1. 理解每一章的基本原理；

2. 如何建模；

3. 分析结果的解读；

4. 针对本章内容的马克威软件操作。

上海天律信息技术有限公司（简称天律公司）十余年来一直秉承软件开发必须与中国的统计和数据挖掘实际相结合的原则，如上海世博会客流量的预测、居民消费结构的研究、大学毕业生就业问题的研究、长三角社会经济发展研究、宏观经济模型的研究、人口问题的研究，都是天律公司社会实践的总结，也是马克威软件在实际应用中的大检验。

自2005年以来，国家统计局开始推出全国统计建模大赛，逢双年为全国统计系统建模大赛，单年为全国大学生统计建模大赛，目的是不断提高广大统计人员的素质，提高中国统计的整体水平。马克威统计分析软件有幸成为全国统计建模大赛推荐的唯一国产软件。为了使参赛选手和广大统计爱好者学好用好马克威软件，天律公司在国家统计局的帮助下特编写此书。

发展是硬道理，天律公司长期以来一直在研究国际上数据分析领域的新思想、新理念、新方法，并将这些新思想、新理念、新方法与中国的实际相结合，旨在推动中国民族统计分析软件的发展，推动中国统计事业的发展。当前互联网已经成为人们生活中不可分割的一部分，天律公司正在与阿里巴巴公司合作，开展基于云计算的分布式数据分析服务，其中包括中小企业贷款评分卡研究，互联网上居民消费行为的研究、互联网云端的挖掘分析研究，等等。天律公司有信心将马克威软件打造成为世界一流的数据分析软件。

本书编写过程中，李鸿琪主持了多个章节的编写，王一、王文青等参加了部分章节的编写。

鉴于时间的紧迫性和我们自身水平的限制，本书中错误或不当之处在所难免，诚恳欢迎同行专家和读者批评指正，并提出宝贵的意见和建议。

黄　晖

2012年8月2日

此为试读，需要完整PDF请访问：www.ertongbook.com

目 录

第一章 上海世博会客流量预测 …… (1)

1.1 世博会简介 …… (1)
1.2 任务 …… (2)
1.3 世博会客流量特点的事前定性分析 …… (2)
1.4 客流量预测技术路线图 …… (3)
1.5 预测结果总结 …… (17)

第二章 扩展线性支出系统(ELES) …… (27)

2.1 扩展线性支出系统及其估计 …… (27)
2.2 扩展线性支出系统的参数估计 …… (28)
2.3 案例分析——上海市区居民消费结构分析 …… (30)
2.4 结论与建议 …… (33)
2.5 马克威软件的相关操作 …… (35)

第三章 二值逻辑回归 …… (38)

3.1 二值逻辑回归的基本原理 …… (38)
3.2 案例分析——上海浦东新区毕业生就业情况调查 …… (41)
3.3 几点建议 …… (47)
3.4 马克威软件的二值逻辑回归操作 …… (48)

第四章 误差修正模型 …… (52)

4.1 误差修正模型(ECM)介绍 …… (52)
4.2 误差修正模型 …… (59)
4.3 案例分析——上海市居民消费实证分析 …… (59)
4.4 马克威软件的误差修正模型操作 …… (65)

第五章　因子分析……（69）

5.1　因子分析基本原理 ……（69）
5.2　案例分析——长三角16城市社会经济综合研究……（73）
5.3　对长三角发展的思考与建议 ……（81）
5.4　马克威软件的因子分析操作 ……（83）

第六章　方差分析模型……（86）

6.1　概述 ……（86）
6.2　案例分析——笔记本电脑价格方差分析模型 ……（90）
6.3　基本假设 ……（90）
6.4　数据采集与模型选定 ……（91）
6.5　描述性分析 ……（92）
6.6　应用模型 ……（98）
6.7　多重比较 ……（102）
6.8　结论与建议 ……（103）
6.9　马克威软件的方差分析模型操作 ……（103）
附件　多重比较结果……（105）

第七章　面板数据(Panel Data)模型 ……（110）

7.1　面板数据介绍 ……（110）
7.2　模型形式的设定检验 ……（117）
7.3　面板数据模型的估计方法 ……（118）
7.4　Hausman检验 ……（118）
7.5　案例分析——中国城镇居民消费研究 ……（119）
7.6　马克威软件的面板数据操作 ……（127）

第八章　偏最小二乘回归……（135）

8.1　偏最小二乘回归原理 ……（135）
8.2　案例分析——三次产业增加值与支出法GDP的关系分析 …（138）
8.3　马克威软件的偏最小二乘回归操作 ……（153）

第九章 PLS 路径模型 …… (155)

9.1 PLS 路径模型介绍 …… (155)
9.2 案例分析——美国顾客满意度指数模型 …… (162)
9.3 马克威软件的路径模型操作 …… (163)

第十章 RBF 神经网络 …… (175)

10.1 案例分析——二手房价格评估 …… (175)
10.2 数据来源 …… (176)
10.3 应用模型 …… (178)
10.4 结论 …… (184)
10.5 马克威软件的 RBF 神经网络操作 …… (184)

第十一章 决策树 …… (196)

11.1 案例分析——电信客户流失预测 …… (196)
11.2 数据来源 …… (197)
11.3 应用模型 …… (197)
11.4 结论 …… (205)
11.5 马克威软件的决策树操作 …… (206)

第十二章 贝叶斯网络 …… (212)

12.1 案例分析——乙型病毒性肝炎诊断 …… (212)
12.2 数据来源 …… (213)
12.3 应用模型 …… (213)
12.4 结论 …… (219)
12.5 马克威软件的贝叶斯网络操作 …… (219)

第十三章 联立方程模型的估计与模拟 …… (223)

13.1 联立方程系统概述 …… (223)
13.2 联立方程系统的基本概念 …… (224)
13.3 联立方程系统的识别 …… (227)
13.4 联立方程系统的估计方法 …… (229)
13.5 联立方程模型的模拟 …… (238)

13.6 案例分析——江城市宏观经济年度模型 …… (242)
13.7 预测 …… (250)
13.8 几点认识 …… (264)
13.9 马克威软件的联立方程操作 …… (266)

第十四章 北京市常住人口预测与分析 …… (285)

14.1 2010年北京市常住人口概况 …… (285)
14.2 预测数据准备 …… (286)
14.3 二大板块的划分 …… (287)
14.4 户籍人口预测方法 …… (288)
14.5 外来常住人口预测 …… (295)
14.6 文化程度预测 …… (299)
14.7 北京市常住人口预测结果 …… (300)
14.8 几点思考 …… (307)

参考文献 …… (312)

第1章 上海世博会客流量预测

摘要：上海市天律信息技术有限公司有幸承担了上海世博会客流量预测的任务，这是一个光荣而又艰巨的任务。根据大量前期数据的调研、分析、研究，最后提出综合动态预测方案，得到了世博局的认可。所谓综合，即不以一种方法来预测，而是采用多种方法进行综合评价，然后预测，所谓动态是在时间上进行滚动预测。上海世博会客流量预测取得了圆满的成功，平均相对误差为5.07%，大大低于爱知世博会的预测（平均相对误差34.9%）和大阪世博会的预测（平均相对误差12.2%）。

1.1 世博会简介

世界博览会（World Expo）又称国际博览会，简称世博会，是一项由主办国政府组织或政府委托有关部门举办的有较大影响和悠久历史的国际性博览活动。参展者向世界各国展示当代的文化、科技和产业上正面影响各种生活范畴的成果。

世博会，是一个富有特色的讲坛。它鼓励人类发挥创造性和主动参与性。把科学性和情感结合起来，将有助于人类发展的新概念、新观念、新技术展现在世人面前。其特点是举办时间长、展出规模大、参展国家多、影响深远。因此，世博会被誉为世界经济、科技、文化的“奥林匹克”盛会。

中国2010年上海世界博览会是第41届世界博览会。于2010年5月1日至10月31日期间，在中国上海市举行。此次世博会也是由中国举办的首届综合类世界博览会，共有246个国家、地区和国际组织参加。上海世博会以“城市，让生活更美好（Better City，Better Life）”为主题，总投资达450亿美元，创造了世界博览会史上最大规模记录，这是中国的骄傲、是上海城市的骄傲，也是上海

市民的骄傲。

上海天律信息技术有限公司有幸成为该次盛会的服务成员，成为世博会运营指挥系统建设成员之一，主要负责数据集成和辅助决策，其中辅助决策系统中最重要的一项为客流量预测。

1.2 任　务

上海世博会组委会、上海世博局的口号是：举办一届“成功、精彩、难忘”的世博会。成功是基础，每天几十万、上百万人的参观，场馆的设施、各项服务设施、安保设施，特别是极端情况下的各项服务是否完善是成功的关键，精彩，一日精彩容易，日日精彩不易，本次世博会就是要将她办成日日精彩的世博会，难忘指一定会有一些特别精彩的人、特别精彩的事使人难忘。

参观人员的有序流动是世博会成功的关键，为此世博局根据场馆的设置和世博的场地面积制定了详细的预警方案：园内人数小于 50 万人为绿灯区；园内人数大于 50 万进入黄灯区；园内人数大于 60 万进入红灯区。绿灯区属于正常参观时间段，黄灯区为二级预警，要在上海地区发出预警，电视、电台发布告知：请市民有序参观，现在参观人数已大于 50 万，请市民不要再前往参观，当园内人数大于 60 万，不但要向上海市民发出告知，还要向长三角地区和全国发出告知：请大家不要涌向上海，现在世博会每日参观人数已达 60 万以上，希望大家安排好参观时间。

根据世博会场馆面积和室外空地面积推算，当参观人数大于 50 万时，参观人员的用餐、用水、卫生设备都会发生一定的拥挤，特别是在炎热的夏天，感觉会不很好，参观的效果也会较差，当参观人数大于 60 万时，参观效果会很差，参观人数过于拥挤也是不安全的主要因素之一。世博局给客流量预测小组下达了任务是：对未来一天、三天和七天的客流量做出准确的预测，平均相对误差控制在 10%以内。可以说客流量的准确预测，是世博会正常运营的保障前提。

1.3 世博会客流量特点的事前定性分析

在制定世博会客流量预测技术路线图前，我们研究了日本大阪世博会、日本爱知世博会的特点和规律，同时也研究了上海科技馆、上海东方明珠等大型场馆的旅游人数的参观特点，在此基础上分析影响世博会客流量因数大致有以下几点：

星期六、星期天、节假日对客流量有明显的影响，特别是星期六影响较大；

此为试读，需要完整PDF请访问：www.ertongbook.com

学生考试、学生暑假期间对客流量有较大的影响，6 月份是上海、长三角中考和高考月份，客流量会走低、7、8 月暑假，客流量会走高；

天气对客流量影响不会很大，但绝对高温和雨季对客流量有一定的影响；

长三角地区，由于邻近上海、经济文化又较发达，所以是上海世博会客流量的主要来源；

客流量的时间分布，5 月，世博会开园月，人们先睹为快的心理驱使会形成一个客流量小高峰，6 月，是中考和高考时期又逢黄梅雨季，客流量会走低，7 月、8 月为暑假期，会形成学生潮小高峰，9 月学生开学客流量会走低，10 月，世博会最后一月，又逢“十一”长假，在双重心理的驱使下，世博客流量将达最高峰。

从整体来看世博会客流量从 5 月至 10 月的分布为高、低、高、低、高的 W 分布态势。

1.4　客流量预测技术路线图

1.4.1　变量选择

可以将世博会客流量分成二部分，一部分是当地居民即在上海居住的居民，影响这部分人的参观因素主要是天气、星期、节假日。另一部分主要是外来人员（非上海常住居民），这部分客流量与铁路、水路、公路和航空进入上海的人数有关，其中团队预约人数是相对确定的未来要参观的人数。表 1 是与世博会客流量有关的自变量表，这些变量有些与模型有关，有些与预警有关。

表 1　与客流量有关的变量

变量名	变量类型	单位	备　注
星期	虚拟变量		虚拟变量赋值
节假日	连续变量	人	人为赋值
退房人数	连续变量	人	旅游局提供
前日预约人数	连续变量	人	世博组委会提供
团队预约人数	连续变量	人	世博组委会提供
宾馆在住人数	连续变量	人	旅游局提供
宾馆入住人数	连续变量	人	旅游局提供
白天最高温度	连续变量		气象局提供
市域客流	连续变量	人	交通局提供当日铁路、水路、公路、航空流入人数

1.4.2 预测技术路线图

根据世博局的要求必须对未来一天、三天、七天的客流量作出预测，如果未来一天会发生大客流量，必须提前一天告知上海市民，做到有序参观，如果预测三天后会发生大客流量，必须提前三天告知长三角居民，请长三角市民暂缓参观世博会，如果预测七天后会出现大客流量，将告知全国国民，暂缓参观世博会。在预测大客流量的同时，世博会有关部门要做好一切安全保卫工作，势必保证世博会的成功。

为保证客流量预测的准确无误(不出现大的误差)，技术路线图必须回答下面二个问题：

- 用什么方法预测？
- 模型的评估标准是什么？

用什么方法预测？统计科学发展到今天，定量预测的模型已经很多，回归模型、指数平滑、ARMA 模型、ARIMA 模型、灰色理论、模糊数学、神经网络等都可以作为预测工具，经过再三斟酌、模拟、权衡，最后选定三种方法为本次选用的预测方法，它们是：逐步回归方法、RBF 神经网络和指数平滑。

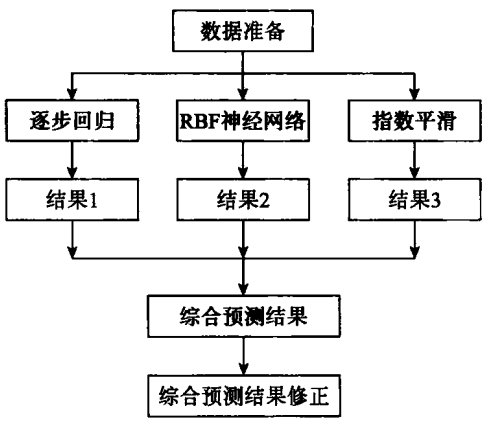

图 1　客流量预测技术路线图

上述流程图的含义是对原始数据分别建立逐步回归模型、RBF 神经网络模型和指数平滑模型，分别分析模型的拟合效果，评估的标准为：

$$\text{平均绝对误差(MAE)} = \frac{1}{h}\sum_{t=T+1}^{T+h} |\hat{y}_t - y_t| \qquad (1-1)$$

$$\text{平均相对误差(MPE)} = \frac{1}{h}\left|\frac{\hat{y}_t - y_t}{y_t}\right| \tag{1-2}$$

根据平均绝对误差的大小和平均相对误差的大小确定这一时期预测方法的优劣程度，从而确定各种方法的权重，综合预测结果为

综合预测结果＝W1＊逐步回归＋W2＊RBF 神经网络＋W3＊指数平滑　(1－3)

其中：W1＋W2＋W3＝1　(1－4)

综合预测结果的修正＝综合预测结果＊星期权数　(1－5)

综合预测结果的修正，由于星期变量在逐步回归中总容易被剔除，而星期变量似乎还是有些规律，所以对星期变量用人工方法设计了权数，见表 2。

表 2　星期权数

星　期	星期权数
星期一	0.75
星期二	0.90
星期三	0.90
星期四	0.90
星期五	0.85
星期六	1.20
星期日	1.20

为什么不选择一种方法，简单明了，而要选择三种方法？我们对日本大阪世博会、爱知世博会大量数据进行了模拟，由于每天的客流量波动较大，有规律又没有规律，一种方法对某一时期拟合可能较好，对另一段时期不一定拟合很好，如单选一种方法预测从整体上来认识会冒较大风险，选择三种方法进行加权综合预测，可以使风险控制在最小。

为什么叫动态综合预测呢？因为从整体来认识，世博会的客流量过程是一个随机过程，整个过程表现为波动性和非线性，而从一段时间来认识这种波动性相对会小一些，所以每次建模的基准数据为建模日期回推 30 天，每过 7 天，回推 30 天建模，直到世博会结束，整个过程是一个动态建模过程。

1.4.3　三种预测方法介绍

1. 逐步回归法

逐步回归法：在回归时，将自变量逐个地引入模型，每引入一个新变量，要对

选入方程的旧变量进行检验，判断其是否显著，剔除不显著的变量，直到既没有新变量选入，又没有旧变量剔除为止。通过逐步回归将显著性变量引入模型，然后依据模型和已知变量对未来 7 天的客流量进行预测和评估。下面以 6 月份数据为例说明逐步回归法的预测过程。

(1)回归建模

根据 6 月份客流量数据，利用逐步回归得到如下回归方程：

$$入园客流 = 134142 * 节假日 + 1.73686 * 团队预约人数$$

模型分析：从回归系数分析表中可以看出节假日和团队预约人数对入园客流量影响比较大，同时模型检验，显著性都小于 0.05，$R^2 = 0.9895$，由于常数项不能通过检验，入园客流可以用节假日和团队预约人数来解释(图 2、表 3、图 3)。

回归系数分析

	回归系数	标准误	标准化的beta	t	显著性
节假日	134,141.9960	50,503.2005	0.3693	2.6561	0.0129
团队预约人数	1.7369	0.3292	0.4957	5.2764	0.0000

图 2　回归系数分析

表 3　拟合值和残差表

行号	原始值	拟合值	残差
001	311080.0000	350025.2549	−38945.2549
002	369595.0000	398805.0234	−29210.0234
003	417516.0000	428120.1944	−10604.1944
004	437030.0000	430430.2208	6599.7792
005	524932.0000	487381.0172	37550.9828
006	417355.0000	405787.2085	11567.7915
007	487873.0000	404880.1520	82992.8480
008	510918.0000	419324.3108	91593.6892
009	413430.0000	418850.1474	−5420.1474
010	391293.0000	432672.5096	−41379.5096
011	403023.0000	434546.5837	−31523.5837
012	424669.0000	456024.6188	−31355.6188
013	417323.0000	389368.6523	27954.3477
014	503182.0000	493219.0249	9962.9751

续表

行号	原始值	拟合值	残差
015	552023.0000	509344.4660	42678.5340
016	379050.0000	415477.9907	−36427.9907
017	394146.0000	423915.2515	−29769.2515
018	414428.0000	451024.1932	−36596.1932
019	429793.0000	487257.7000	−57464.7000
020	361227.0000	394488.9214	−33261.9214
021	415091.0000	430596.1304	−15505.1304
022	409818.0000	446738.9401	−36920.9401
023	404127.0000	447121.0498	−42994.0498
024	447153.0000	459722.3980	−12569.3980
025	480931.0000	464641.1911	16289.8089
026	553480.0000	486934.6437	66545.3563
027	486766.0000	405488.4683	81277.5317
028	458359.0000	446941.7384	11417.2616
029	452572.0000	462026.7992	−9454.7992
030	427936.0000	422011.2362	5924.7638

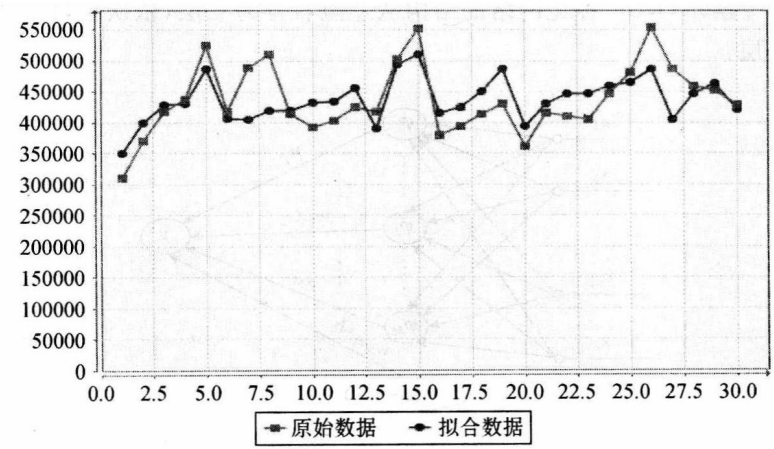

图 3　入园客流拟合图

(2)回归预测

未来 7 天入园客流量预测,根据已知模型,将未来 7 天的外生变量值输入,这里是节假日变量和团队预约人数,得到未来 7 天入园客流的预测值表。

表 4　未来 7 天入园客流预测值

行号	入园客流预测值
1	421308.2217
2	453396.7467
3	470635.9309
4	435694.2349
5	432487.5731
6	415522.3199
7	425306.0634

2. RBF 神经网络

RBF 神经网络是一种三层前向网络,通过输入层空间到隐含层空间的非线性变换以及隐含层空间到输出层空间的线性变换,实现输入层空间到输出层空间的映射。这两个层间变换参数的学习可以分别进行,使得 RBF 神经网络的学习速度较快且可避免局部极小问题。

如图 4 所示,RBF 神经网络的结构从左至右分为三层,依次是输入层、隐含层和输出层:

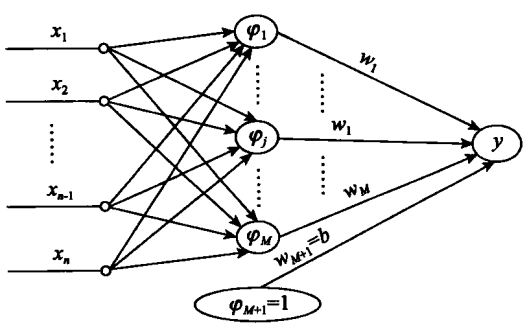

图 4　RBF 神经网络结构

此为试读,需要完整PDF请访问: www.ertongbook.com