

教育部人文社会科学研究规划基金项目（项目批准号：14YJAZH055）

基于大数据挖掘的 服刑人员再犯罪预测

马国富 著



中国政法大学出版社

基于大数据挖掘的 服刑人员再犯罪预测

马国富 著



中国政法大学出版社

2018 · 北京

- 声 明
1. 版权所有，侵权必究。
 2. 如有缺页、倒装问题，由出版社负责退换。

图书在版编目（CIP）数据

基于大数据挖掘的服刑人员再犯罪预测/马国富著. —北京：中国政法大学出版社，2018. 11

ISBN 978-7-5620-8730-4

I. ①基… II. ①马… III. ①信息技术—应用—监狱—管理—研究 IV. ①D916. 7-39

中国版本图书馆 CIP 数据核字(2018)第 267787 号

- 出 版 者 中国政法大学出版社
- 地 址 北京市海淀区西土城路 25 号
- 邮寄地址 北京 100088 信箱 8034 分箱 邮编 100088
- 网 址 <http://www.cuplpress.com> (网络买名：中国政法大学出版社)
- 电 话 010-58908285 (总编室) 58908433 (编辑部) 58908334 (邮购部)
- 承 印 固安华明印业有限公司
- 开 本 880mm × 1230mm 1/32
- 印 张 12.5
- 字 数 356 千字
- 版 次 2018 年 11 月第 1 版
- 印 次 2018 年 11 月第 1 次印刷
- 定 价 46.00



监狱作为国家的刑罚执行机关，是维护社会稳定的重要力量，监管安全是监狱各项工作的基础，也是实现刑罚执行目的的前提条件。目前，监狱提出了各种管理方法，制定了各种管理制度来规范监管，以确保监管安全。近年来，按照国家、司法部和各省的有关部署，经过各级司法机关的共同努力，监狱信息化建设工作已取得了很大的进展，但各地监狱信息化应用的总体水平仍然相对较低，信息技术在监管安全中的应用有待进一步提升。随着云计算、物联网、智能化视频监控等新型 IT 技术在监狱中的深入应用，监狱网络、信息资源库、应用软件、应用服务器、视频监控系统、无线传感器网络、基于无线定位的电子腕带和 RFID 等组成的物联网智能安防监控等系统所产生的数据呈爆炸式增长，并且数据量从线性级到指数级增长。数据已经成为一种新的资产，而大数据将产生新的价值，监狱系统正面临着“大数据”“大系统”的管理和维护问题。通过到有代表性的监狱进行调研，我们发现，各地区的监狱信息化建设取得了很大进展。监狱建立了大数据中心，实现对各业务数据的整合和集中存储，但是对海量数据的分析、挖掘还处于初级阶段，监狱系统基本上实现了业务的数据化，监狱急需实现数据的业务化。以大数据为引领，围绕“政

治改造、监管改造、教育改造、文化改造、劳动改造”新格局，将物联网、云计算、移动互联等信息技术与监管改造工作深度融合，对监狱各类信息进行实时、精确、全面地感知、整合和分析，全方位支撑监狱民警执法、风险管控、教育改造、队伍建设、综合保障等方面智慧化发展，实现监狱管理精细化、指挥调度立体化、安全防控精准化、刑罚执行智能化、教育矫治科学化、综合办公无纸化，助推监管改造工作在新时代实现新发展。

在对服刑人员再犯罪概念进行精确界定的基础上，利用大数据挖掘技术从监狱信息化资源库、安防监控等系统及服刑人员的日常行为中收集服刑人员的相关数据，建立监狱大数据收集的规范化流程，并对收集的数据进行数据清洗、数据集成、隐私数据脱敏、数据变换和数据规约等数据预处理操作；从预处理后的数据中提取和选择涉及服刑人员危险性和再犯罪的相关特征，建立训练集和测试集，然后基于不同监狱内应用场景和数据类型使用聚类、关联、分类和回归算法进行交叉验证训练数据，建立基于大数据的服刑人员再犯罪预测模型，获取规律性知识和洞察来对监狱服刑人员的再犯罪进行模式识别和预测，服刑人员再犯罪的识别与预测将日益基于数据分析做出，而不是像过去更多凭借经验和直觉。

本书由马国富编写大纲，并撰写各章节内容。课题组成员王子贤、马胜利、刘恒志曾到浙江省乔司监狱、浙江省女子监狱、浙江省金华监狱进行实地调研、座谈，并和浙江警官职业学院相关教师就服刑人员再犯罪预测进行学术交流，之后他们给撰写本书提供了很好的建议；课题组成员汪玉红收集服刑人员改造质量、服刑人员计分考核等相关资料，并对服刑人员计分考核在监狱中存在的问题进行了研究，为本书的撰写提供了很好的参考。

本书主要是利用大数据技术对服刑人员再犯罪进行预测，但是大数据技术也可运用于监狱的五大改造等方面，为构建“数字法治、智慧司法”信息化体系，形成“大平台共享、大系统共治、大数据慧治”的信息化新格局提供帮助和参考。在编写过程中参考了大量著

作、论文和电子文献，并到浙江、江西、北京、河北、江苏、广西、云南、内蒙古、贵州等地监狱进行了实地调研和座谈，在此向所有文献资料的作者及提供调研帮助的监狱表示由衷的感谢。另受到时间、经费等多方面因素的制约，书中存在的不妥之处还请同行专家及学者批评指正。

本书是2014年度教育部人文社会科学研究规划基金项目（项目编号：14YJAZH055）的最终成果，并得到相关资助，在此表示感谢。

马国富

2018年11月7日

◇ 前 言	1
◇ 第 1 章 服刑人员再犯罪预测	1
1.1 服刑人员再犯罪概述	1
1.1.1 再犯罪概念的界定	1
1.1.2 再犯罪危险的界定	6
1.1.3 服刑人员再犯罪现状	8
1.1.4 服刑人员再犯罪原因分析	10
1.1.5 服刑人员再犯罪防控新机制	13
1.2 服刑人员危险性评估	16
1.2.1 服刑人员危险性评估定义	17
1.2.2 国内服刑人员危险性评估	18
1.2.3 国外服刑人员危险性评估	21
1.3 服刑人员再犯罪预测	29

1.3.1 国外服刑人员再犯罪预测	30
1.3.2 国内服刑人员再犯罪预测研究	34
◇ 第2章 服刑人员再犯罪数据挖掘流程	37
2.1 服刑人员再犯罪数据挖掘概述	37
2.1.1 数据挖掘的定义	37
2.1.2 数据挖掘的对象	38
2.2 服刑人员再犯罪数据挖掘目标	43
2.2.1 数据挖掘目标的定义	43
2.2.2 数据挖掘方法	44
2.2.3 数据挖掘目标的团队构成	48
2.3 国外跨行业数据挖掘标准过程 (CRISP-DM)	50
2.3.1 商业理解阶段	51
2.3.2 数据理解阶段	54
2.3.3 数据准备阶段	56
2.3.4 建模阶段	59
2.3.5 模型评估阶段	61
2.3.6 模型发布阶段	63
2.4 国内数据挖掘预测流程	66
2.4.1 定义问题	67
2.4.2 准备数据	68
2.4.3 选择模型	71
2.4.4 构建模型	73
2.4.5 评估与优化模型	81
2.4.6 部署模型	99

◇ 第3章 服刑人员再犯罪预测数据准备	101
3.1 服刑人员再犯罪数据的收集	101
3.1.1 认识数据	101
3.1.2 大数据	111
3.1.3 服刑人员再犯罪预测的数据源	127
3.1.4 大数据时代下的服刑人员再犯罪数据抽样	133
3.2 读取数据	136
3.2.1 从 CSV 文件中读取数据	137
3.2.2 从 Microsoft Excel 文件中读取数据	140
3.2.3 从 XML 文件中读取数据	141
3.2.4 从 JSON 数据源读取数据	143
3.2.5 从数据库文件中读取数据	143
3.3 服刑人员再犯罪数据质量分析	144
3.3.1 数据质量分析指标	145
3.3.2 数据质量基本理论	146
3.3.3 数据质量分析	154
◇ 第4章 数据预处理	158
4.1 数据预处理概述	158
4.2 数据清洗	160
4.2.1 缺失数据处理	161
4.2.2 冗余数据处理	165
4.2.3 噪声数据处理	167
4.3 数据集成	171
4.3.1 数据集成基本类型	171

4.3.2 数据集成存在的问题	172
4.3.3 数据集成方法	174
4.4 隐私数据脱敏	177
4.4.1 隐私数据泄露类型	178
4.4.2 隐私数据脱敏概述	179
4.4.3 服刑人员隐私数据脱敏算法	182
4.4.4 数据脱敏方法	187
4.4.5 大数据脱敏平台	190
4.5 数据变换	191
4.5.1 简单函数变换	191
4.5.2 数据标准化	192
4.5.3 特征离散化	196
4.6 数据规约	205
4.6.1 维数灾难与过拟合	206
4.6.2 维规约	213
4.6.3 数值规约	220
◇ 第5章 服刑人员再犯罪数据挖掘建模	224
5.1 服刑人员再犯罪数据挖掘概述	224
5.2 关联规则挖掘	227
5.2.1 关联规则挖掘概述	227
5.2.2 Apriori 算法	232
5.2.3 FP-Growth 算法	237
5.3 回归分析方法	245
5.3.1 回归分析方法概述	246

5.3.2 线性回归	248
5.3.3 非线性回归	252
5.3.4 回归方法检验	254
5.4 分类方法	254
5.4.1 分类方法概述	255
5.4.2 逻辑回归	256
5.4.3 K 近邻分类	260
5.4.4 贝叶斯分类	265
5.4.5 支持向量机	269
5.4.6 决策树分类	285
5.4.7 神经网络分类	300
5.5 集成学习	312
5.5.1 集成学习概述	312
5.5.2 Bagging	316
5.5.3 Boosting	320
5.5.4 Stacking	324
5.6 聚类分析方法	329
5.6.1 聚类分析方法概述	329
5.6.2 聚类分析方法	330
5.6.3 相似度的度量	331
5.6.4 K-means 聚类	333
5.6.5 K-medoids 聚类算法	338
5.6.6 聚类分析总结	340
5.7 基于离群点检测的服刑人员安全监管改造分析	341
5.7.1 离群点概述	341

5.7.2 离群点类型	342
5.7.3 离群点检测方法	343
◇ 第6章 基于大数据挖掘的服刑人员再犯罪预测	349
6.1 基于大数据的服刑人员危险性预测研究	349
6.1.1 监狱监管改造安全的现状	350
6.1.2 服刑人员再犯罪预测与危险性评估	351
6.1.3 监狱大数据分享中的隐私保护	354
6.1.4 基于大数据的服刑人员危险性识别与预测	357
6.1.5 未来展望	365
6.2 机器学习模型在预测服刑人员再犯罪危险性中的效用	366
6.2.1 服刑人员危险性评估现状	366
6.2.2 机器学习模型	369
6.2.3 机器学习模型数据源	374
6.2.4 经典机器学习模型在预测服刑人员再犯罪危险性 中的效用	375
6.2.5 未来展望	383



服刑人员再犯罪预测

1.1 服刑人员再犯罪概述

近年来，再犯罪现象日益突出，犯罪主体呈现职业化、系列化、团伙化，再犯罪手段更为成熟隐蔽、反侦察能力更为强大、再犯罪造成的社会危害性也更加严重，给社会的安全稳定带来了巨大的负面影响。目前，监狱信息化逐步实现了从数量扩展到质量提升，云计算、物联网、大数据等新型 IT 在监狱中的深度应用实现了“人防、物防、技防”的深度融合。利用大数据对涉及服刑人员的所有数据进行收集、分析，建立基于大数据挖掘的服刑人员再犯罪预测模型，从而实现对服刑人员再犯罪的识别、预测和重点监控，重塑监管改造工作的新范式。利用大数据技术可实现对服刑人员的个性化改造和精准化预测，增强对服刑人员教育改造的预见性、针对性和实效性，提升监狱科学管理的信度和效度，为新时期监狱从底线安全观向治本安全观转变提供保障。

1.1.1 再犯罪概念的界定

科学、准确地界定再犯罪的内涵与外延，是研究再犯罪现象的基本前提。反思与重构再犯罪的概念，首先需要对已有的概念界定做出系统而全面的梳理。界定再犯罪的目的是为了构建再犯罪的预测、预警机制，有效预防、减少和控制再犯罪的发生，实现从“底线安全

观”向“治本安全观”的转变，切实提高服刑人员的改造质量。再犯罪又称为重新犯罪，英语一般用“recidivism”，《高阶英汉双解词典》解释为屡教不改的服刑人员、惯犯、累犯。^{〔1〕}将累犯定义为因为先前的犯罪而被宣判有罪后由于犯了新罪再次被定罪的人^{〔2〕}。在刑法学、犯罪学、监狱行刑学等学科分类的基础上，我们对现有的具有代表性的再犯罪概念归纳整理如下。

1. 基于刑法学角度界定的再犯罪

基于刑法学的角度对再犯罪进行界定的主要目的是为司法实践中的定罪量刑提供参考，张广智等学者认为^{〔3〕}：再犯罪又称重新犯罪，是指由于触犯刑法被判刑，在服刑结束或被释放回归社会后又因犯罪被判刑的行为。隗甫杰等学者认为^{〔4〕}：再犯罪是指犯罪前科人员“二次犯罪”乃至累次犯罪的重复犯罪。综上所述，广义的再犯罪既包括行为人接受处罚后的狱外犯罪，又包括服刑人员在监狱内再犯罪；而狭义的再犯罪主要是指行为人刑满释放后的再犯罪。

2. 基于犯罪学角度界定的再犯罪

基于犯罪学的角度对再犯罪进行界定的主要目的是预防和控制再犯罪，周路等学者认为^{〔5〕}：解除劳动教养或刑罚执行完毕的人员，在任何时间再实施刑法规定的犯罪行为并接受刑罚处罚的为重新犯罪；郑祥认为^{〔6〕}：重新犯罪是指有犯罪前科，即在此次被判刑前，

〔1〕 Georgia Zara, David P. Farrington, *Criminal Recidivism Explanation, Prediction and Prevention*, New York: Routledge, 2016.

〔2〕 笔者译。文献原文为：recidivism is the official criminal involvement (based on criminal records) of a person who, after having been convicted for a previous offence, commits a new crime for which they incur another conviction. p. 5.

〔3〕 参见张广智、向静：“对当前刑满释放人员再犯罪的调查分析”，载《法制与社会》2010年第22期。

〔4〕 隗甫杰、梁兵、刘伟：“论当前再犯罪特点及刑侦工作对策”，载《北京警察学院学报》2015年第1期。

〔5〕 参见周路、刘文成、王志强：《当代实证犯罪学新编——犯罪规律研究》，人民法院出版社2004年版。

〔6〕 郑祥：“防治重新犯罪与构建和谐社会——重新犯罪现状与对策的实证研究”，载《吉林公安高等专科学校学报》2007年第6期。

曾经因为犯罪行为受到过劳教或司法处分判刑的行为。综上所述,广义的再犯罪既包括刑满释放人员的再犯罪和正在服刑期间的服刑人员犯罪,又包括经公安机关处理的正在进行劳动教养或已经解除劳动教养的人员的犯罪行为;而狭义的重新犯罪是指已经解除劳动教养和刑满释放人员实施的再犯罪行为。相比较于刑法学,犯罪学不仅包括曾经判刑入狱的服刑人员也包括被劳动教养的人员,2013年11月,十八届三中全会通过《中共中央关于全面深化改革若干重大问题的决定》,废止了劳动教养制度。

3. 基于监狱行刑学角度界定的再犯罪

基于犯罪学的角度对再犯罪进行界定的主要目的是:使用刑法规定的标准计量再犯罪率,衡量服刑人员改造质量,提高服刑人员改造质量。由于不同时期的刑法对重新犯罪内容的规定有所不同,导致重新犯罪的概念也不同。白正春学者等认为〔1〕:重新犯罪是指触犯刑事法律并受到刑罚处罚,回归社会后又重新故意实施犯罪活动,依法应当追究其刑事责任的行为。

江华锋〔2〕综合刑法学、犯罪学、社会学等学科的观点将重新犯罪定义为行为主体受过刑罚之宣告后,在再社会化期间或结束后,再次有意识实施侵犯其他主体合法权益,依法应当被追究法律责任并需要采取社会防范和控制措施的犯罪行为。这一界定主要包含了重新犯罪的性质、主体、主观条件、客观条件、场域、目的等六大因素。重新犯罪的性质因素是指:重新犯罪必须是具有前后两次或两次以上反社会性、刑事违法性、社会危害性的独立的犯罪行为,是犯罪的刑事违法性和社会危害性的统一。重新犯罪的主体因素是指:受过刑罚宣告之后再犯罪的自然人,不包括法人、非法人单位等。受过刑罚之宣告的人包括:在监狱、少管所、看守所服刑人员,被判处管制、宣告缓刑、假释或暂予监外执行的社区服刑人员,被单独判处有期徒刑、没收

〔1〕 白正春、杨冰川:“论和谐社会视野下重新犯罪问题及对策”,载《南方论刊》2010年第12期。

〔2〕 江华锋:“我国重新犯罪概念的再界定”,载《学海》2017年第3期。

财产、剥夺政治权利人员，免于刑事处罚人员，附条件不起诉对象以及被收容教养对象。重新犯罪的主观条件因素包括罪过形式和人格因素两部分。罪过形式主要指判断犯罪的主观情况，无论首次犯罪是否故意，只要第二次及以后犯罪为故意犯罪，就属于重新犯罪；人格因素主要指人身危险性。重新犯罪的客观条件因素包括罪次条件、时间条件两个部分。其中，罪次条件是指行为人进行了前后两次或两次以上的独立犯罪，即：不论是在前罪的刑罚执行完毕或者赦免以后，抑或发生在刑罚执行期间，只要再进行犯罪就属于重新犯罪；时间条件是指确定初犯与再犯之间的时间间隔，前罪和后罪之间没有时间间隔限制，只要是前罪刑罚宣告之后的任何时间的再犯罪，都属于重新犯罪。重新犯罪的场域因素是指重新犯罪的空间场所，可能是监狱、少管所及看守所，也可能是社区以及其他社会场所。重新犯罪的目的因素是指对重新犯罪有一个统一、准确、全面的认识，进而形成统一衡量服刑人员改造质量的计量标准，从而构建再犯罪预测、预警及控制机制，有效预防、减少和控制再犯罪的发生。

曾贇从规范意义和统计意义两个角度对再犯罪的定义进行了界定〔1〕。规范意义上的定义主要是依据法律辞典的解释，主要分为客观事实和主观心理两个层面。以《牛津法律大辞典》为代表的一类辞典从再犯罪事实的角度进行界定，将再犯罪定义为被释放的囚犯再次犯罪，并被重新定罪。从上我们可以得出：如果某人犯罪被判刑入狱后，因再次犯罪被定罪，不论该罪被处以何种刑罚，则均被认为属于再犯罪。以《布莱克斯通法律辞典》为代表的一类辞典从再犯罪的心理结构角度进行界定，将再犯罪定义为一种重新陷入犯罪活动或犯罪行为的倾向。从上我们可以得出：一个曾经犯罪被判刑入狱的人再次陷入犯罪倾向的被认为是再犯罪，显然从法律角度上来看是站不住脚的，这个很大程度上属于再犯罪预防的范畴，可以利用大数据技术来实施再犯罪的预测和预警，这是监狱现在及未来一段时间内亟需开展

〔1〕 曾贇：“论再犯罪危险的审查判断标准”，载《清华法学》2012年第1期。

的工作。再犯罪统计意义上的定义主要是指相关机构发布的通知和文件上的指标。新西兰矫正局的调查统计报告〔1〕中将再犯罪定义为：因犯罪被判刑入狱或社区矫正释放后再次犯被判刑入狱或社区矫正的犯罪行为。我国政法系统发布的通知和公文中一般把再犯罪称为重新犯罪，最早出现于1950年3月13日司法部发布的《关于假释人犯重新犯罪如何撤销假释问题的批复》，之后最高法、最高检、公安部、司法部陆续出台的公文都以重新犯罪进行界定〔2〕。从上我们可以看出我国政法机关采用刑法中关于重新犯罪的规定来界定再犯罪。1985年1月司法部在其发布的《关于刑满释放、解除劳教人员重新犯罪、违法问题的几点意见》中将再犯罪界定为：原犯普通罪的，刑满释放或赦免以后，在三年以内再犯应判处刑罚的为重新犯罪；原犯反革命罪刑罚执行完毕或者赦免以后，在任何时候再犯反革命罪的，或者三年以内再犯其他普通刑事罪而被判处刑罚的都是重新犯罪。此后，我国司法部门未发布关于再犯罪调查的具体意见。

法律规定什么样的行为是犯罪，会随着社会的发展而变化，在同一国家，有的行为在某一历史时期被认为是犯罪，但在另外一个时期就不被认为是犯罪。所以，应当历史地看待再犯罪概念中的犯罪界定。综合已有再犯罪概念的各种观点，从客观上看，所犯的前罪与后罪都必须构成犯罪，本书所指的犯罪主要指的是违反我国当前刑法规定的犯罪，而再犯者既包括行为人接受处罚后的狱外再犯罪，又包括服刑人员在监狱内的再犯罪。

〔1〕 Reconviction Patterns of Released Prisoners, "A 36-months Follow-up Analysis", Arul Nadesu Strategic Analysis Team Policy Development Department of Corrections, 2007.

〔2〕 1955年9月29日公安部下发的《关于刑满留场就业人员逃跑及重新犯罪的处理问题的批复》，1956年7月5日最高人民检察院在《关于处理劳动改造队加、减刑等法律程序的通知》，1956年9月4日司法部在《关于劳改犯刑期届满前或届满后留场重新犯罪如何确定其罪名的函》，1963年7月29日最高人民法院、最高人民检察院、公安部联合下发的《关于监外执行的罪犯重新犯罪是否履行逮捕手续的批复》，1963年11月7日公安部《关于严防刑满释放分子重新犯罪的通知》，1979年9月24日最高人民法院下发的《关于留场（厂）就业人员重新犯罪后在劳改机关禁闭审查日期应否折抵刑期的批复》等文件中都使用了重新犯罪这一概念。