

一种基于数据挖掘技术的入侵检测方法的设计

刘 犇¹, 毛燕琴², 沈苏彬²

(1. 南京邮电大学 物联网学院, 江苏 南京 210003;

2. 南京邮电大学 计算机学院, 江苏 南京 210003)

摘 要:数据挖掘技术可应用于入侵检测方法中,其中典型的聚类算法 k-means 是一种高效的、可用于分类入侵检测数据的轻量级算法,但该算法存在收敛于局部最优解的问题。针对此问题,提出将遗传算法与 k-means 聚类算法相结合的 GCAH (Genetic and Clustering Analysis Hybrid) 入侵检测方法,对数据进行分析 and 检测,可避免产生聚类算法收敛于局部最优解的问题。利用 KDD cup 网络流量集作为输入数据对 GCAH 入侵检测方法进行实验测试。实验结果表明 GCAH 方法能有效提高检测率、降低误报率,达到预期效果。

关键词:入侵检测;数据挖掘;聚类分析;遗传算法

中图分类号:TP309

文献标识码:A

文章编号:1673-629X(2011)08-0241-05

An Intrusion Detection Method Using Data Mining Technology

LIU Ben¹, MAO Yan-qin², SHEN Su-bin²

(1. College of Internet of Things, Nanjing University of Post and Telecommunications, Nanjing 210003, China;

2. College of Computer, Nanjing University of Post and Telecommunications, Nanjing 210003, China)

Abstract: Data mining can be used in intrusion detection. K-means is an efficient, representative lightweight algorithm which can be used in intrusion detection. However, it has the problem that converges to local optimization. To solve this problem an intrusion detection method called GCAH (Genetic and Clustering Analysis Hybrid) is brought forward. GCAH is a method that combines clustering algorithm with genetic algorithm to analyse data of intrusion detection and solve the problem that converges to local optimization. Testing experiment uses KDD cup data which has lots of network data messages as the input data to prove that GCAH has a high detection rate, low false alarm rate and achieves the goal.

Key words: intrusion detection; data mining; cluster analysis; genetic algorithm

0 引 言

现今的入侵检测技术存在种种不足,基于误用检测的技术只能对已知类型的攻击进行检测^[1],而基于异常检测的技术由于可以检测出未知类型的攻击而受到广泛关注和研究。数据挖掘技术作为一种有效地从大量数据中提取所需信息的方法可以运用于异常检测中,相比于其它入侵检测方法具有扩展性和自适应性较好等优点。目前也有一些从数据挖掘角度对入侵检测的方法研究^[2],但是直接将数据挖掘算法应用在入侵检测中会引入算法原有的一些固有缺陷,如收敛于局部最优解等,从而影响了入侵检测的效果。所以,文中从数据挖掘角度研究网络入侵检测方法,将遗传算

法与聚类算法相结合,提出 GCAH (Genetic and Clustering Analysis Hybrid) 入侵检测方法对数据进行分析 and 检测,从而解决聚类算法收敛于局部最优解的固有缺陷。最后进行了实验和测试,证明 GCAH 入侵检测方法能避免 k-means 聚类算法本身收敛于局部最优解的问题,从而提高检测率、降低误报率,提高效率,达到了预期的效果。

1 入侵检测技术

在当今的企业应用环境中,安全是所有网络面临的大问题。

入侵检测系统 (Intrusion Detection System, IDS) 监控网络及系统的状态、行为和使用,检测用户的越权使用和入侵者利用系统安全缺陷对系统进行入侵的企图^[3]。它已经发展成为安全网络体系中的重要组件。目前市场中也有很多入侵检测工具,包括商品化的和开放源码形式的,可以用来检测网络中存在的不同类型的安全问题。

收稿日期:2011-01-13;修回日期:2011-04-17

基金项目:国家高技术(863)计划项目(2006AA01Z208);江苏省科技支撑计划项目(BE2009157)

作者简介:刘 犇(1985-),男,江苏南京人,硕士研究生,研究方向为计算机网络应用技术;沈苏彬,研究员,博士研究生导师,研究方向为计算机网络、下一代电信网及网络安全。

(1) 误用检测技术。

误用检测对那些试图以非标准手段使用系统的安全事件进行鉴别。IDS 中存储着已知的入侵行为描述,它以此为根据比较系统的行为。

(2) 异常检测技术。

这种检测被用于应用程序层次监控用户的行为。异常检测 IDS 从用户的系统行为中收集一组数据。这一数据集被视为“正常调用”。如果用户偏离了正常调用模式,就会产生警报。

传统的入侵检测技术仍然存在种种不足^[4]。基于误用检测的入侵检测系统存在无法检测未知类型的攻击,需要人工不断更新入侵行为特征库的问题^[5],并且不断增长的特征库也会降低入侵检测的效率。而基于异常检测的入侵检测系统可以从一定程度上检测未知类型的攻击,不依赖于入侵检测的特征库。数据挖掘技术^[6,7]可以有效地从大量数据中提取所需的信息,可以运用于异常检测中^[8],但是需要研究有效的基于数据挖掘技术的入侵检测方法。

2 数据挖掘技术

数据挖掘指从大量的、有噪声的原始数据中提取出隐含在其中的、人们事先不知道的、有用的信息和知识的过程^[9,10]。

数据挖掘的过程如图 1 所示。

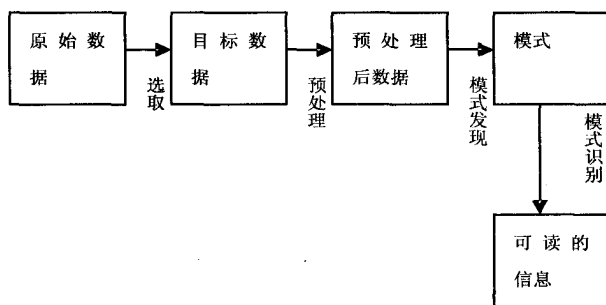


图 1 数据挖掘过程

随着数据挖掘技术的不断发展,现今几种主要的数据挖掘技术如下:

分类分析:利用分类器划分数据。从数据中选出已分类的数据集,依此建立分类模型,再对原数据集中未分类的数据进行分类。

关联分析:找出数据库中隐藏的关联网。发掘大量数据中的相互关联及联系。

序列分析:根据时间序列发掘数据在时间上的联系。

聚类分析:这是一种无监督的学习方法,可以用于对未标记的数据进行分析。依靠一定的相似度准则将数据进行分类,并使得同类之间相似度尽量高而异类之间相似度尽量低。聚类算法中常用的方法有划分方

法、层次方法、密度方法、网格方法、模糊聚类等。

由于入侵检测的数据都是未经过处理和标明的原始数据,所以聚类分析的无监督特性比较适用于入侵检测。同时,由于入侵检测数据量非常大,所以应该采用轻量级的聚类算法。

3 一种基于聚类分析和遗传算法的入侵检测方法

3.1 现有的聚类算法及相关分析

聚类算法中,k-means 算法是一种典型的轻量级划分算法,目前也有一些它在入侵检测方面的应用研究^[2],它的算法的主要步骤如图 2 所示。

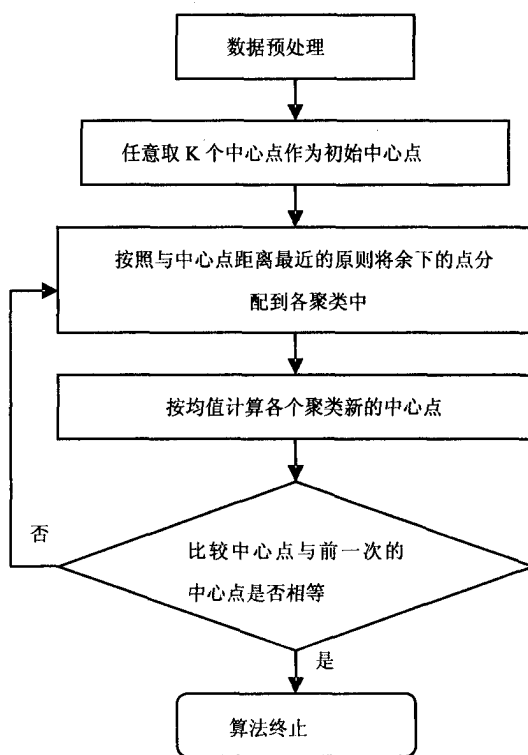


图 2 k-means 聚类算法流程

k-means 算法本身是一个比较高效的聚类算法。但是,k-means 算法本身有其缺陷性:它的局部搜索能力较好,但缺乏良好的全局搜索能力。对于某些初始中心点所求得的聚类可能收敛于局部最优解。即在寻找最佳聚类的过程中,当搜索到局部的极值附近时,算法会误将此极值作为全局的最优解,收敛于局部最优解,算法过早的结束。这就会导致算法所产生的聚类质量不高。

对于入侵检测系统而言,聚类算法的作用就是将待检测的数据连接记录进行分类的过程。

聚类算法可以将待检测的数据连接记录划分为多个聚类:

(1)对于正常的和异常的数据连接,可以根据各个正常的数据连接之间的相似度较高,正常数据连接

和异常数据连接之间的相似度较低的特点将其分为两类。

(2)而在异常数据连接中,又可以根据同一攻击类型之间的相似度较高而不同的攻击类型之间的相似度较低进一步划分为多个聚类。

也就是说,聚类算法所产生的聚类对应了正常数据连接的集合和各类异常数据连接的集合。聚类本身的质量好坏决定了入侵检测结果中各类数据集合分类的正确率,也就决定了入侵检测的检测率和误报率。所以 k-means 算法的局部收敛问题会降低入侵检测的检测率,提高误报率。

3.2 基于聚类分析和遗传算法的入侵检测方法

GCAH

由于 k-means 算法的局部收敛问题会使聚类质量变坏,从而降低入侵检测的检测率。因此,文中引入了遗传算法^[11]来对此问题进行解决,提出了一种将遗传算法和聚类算法相混合的入侵检测方法 GCAH(Genetic and Clustering Analysis Hybrid)。

文中主要是利用遗传算法对 k-means 算法流程中的任意取初始中心点这个部分进行改进,另外根据入侵检测的数据各个属性之间的数值范围变化较大的情况重新设计了数据的预处理模块。

(1)遗传算法是一种自适应全局优化搜索算法。它具有良好的全局搜索能力,这一特点可以弥补 k-means 算法的缺点。所以文中利用遗传算法对数据进行全局搜索寻找合适的 k-means 聚类算法的初始中心点,对 k-means 算法的搜索过程进行修正,而因为遗传算法优良的全局搜索能力遗传算法搜索得到的初始中心点可以最大程度地避开局部的最优解,由这样的初始中心点所得到的聚类结果也会避免发生局部收敛。从而提高聚类质量,提高入侵检测的检测率,降低误报率。

(2)入侵检测是对每一个数据报进行分析,在数据报中有多个数据项,在数据挖掘中将其看作连接记录的属性,由于这些属性值有数值型的、字符型的,所以为了提高入侵检测的效率,将它们统一数值化为数值型的数据,对于原有的字符型数据采用逐个编号的方式数值化。

由于数据连接记录的各个属性值差距很大,例如连接时间、传送的字节数可能从 0 到数万不等,而连接正常或者错误的标识可能不超过 10 种,如果不加处理则会造成属性范围较小的属性的作用在后面的聚类算法中被过分地低估,这会影响入侵检测的准确率。所以在文中的设计中,首先将输入数据中各个属性值归一化为范围在 $[0,1]$ 之间的精度为 6 位小数的数值,使得数据记录中不同属性的作用在后面的聚类算法中

都能得以体现。

由于对入侵检测环境下的数据的分析采用原有的 k-means 算法的收敛条件所得到的聚类效果并不好,文中采用的聚类收敛条件是将前一次聚类得到的结果与本次聚类得到的结果比较,仅当各聚类中的元素完全一致时才判断算法收敛并结束算法。这样最后所得的聚类更为精确,聚类内部的相似度更高、外部的相似度更低。

GCAH 入侵检测方法总体流程如图 3 所示。

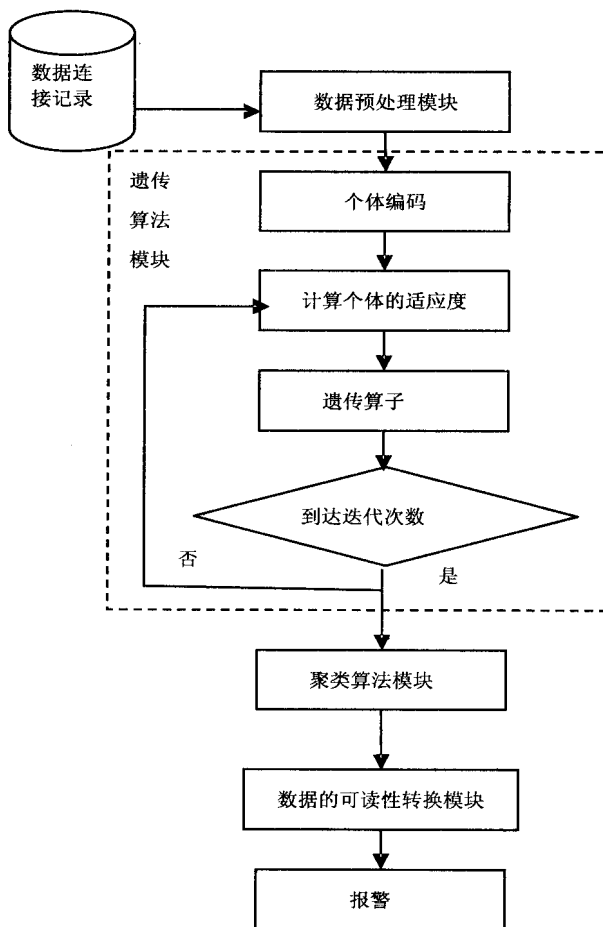


图3 GCAH入侵检测方法总体流程

图3中主要的模块说明如下:

(1)数据的预处理模块:由于数据连接记录的各个属性值差距很大,所以需要对其进行数值化和归一化。

(2)遗传算法模块:文中采用遗传算法对数据先进行处理,利用遗传算法良好的全局搜索能力找到聚类算法模块所需要的合适的初始中心点,避免由于中心点的取值不当导致 k-means 算法收敛于局部最优解。

在遗传算法的编码方法设计中,由于采用浮点数编码可能导致在交叉、变异部分的运算中有可能破坏原有的较好的基因型,所以文中采用了二进制的编码方法,以便于保护已经产生的基因型。

在遗传算子方面,选择算子采用比例选择算法,交叉算子采用单点交叉算子,变异算子采用基本位变异算子。

由遗传算法搜索得到合理的中心点,作为聚类算法的输入参数。这样可以避免聚类算法陷入局部最优解。

(3) 聚类算法模块:根据输入的中心点,对待检测数据进行聚类分析和归类,使得同一聚类内部的相似度尽量高,不同聚类之间的相似度尽量低,得到最终的聚类结果。

(4) 数据的可读性转换:由聚类算法模块得到的结果,判断待检测数据中的攻击数据流量和正常数据流量,以达到入侵检测的目的。

(5) 报警:根据入侵检测的结果向系统管理员报警。

4 实验测试和结果

采用 KDD cup 99 数据集对 GCAH 入侵检测方法的效果进行测试。KDD cup 数据集来源于美国国防部高级规划署(DARPA)在 MIT 林肯实验室进行的一项入侵检测评估项目。林肯实验室建立了一个模拟的网络环境,收集了数周时间的 TCPdump 连接数据和主机审计数据,模拟了各种不同的攻击手段,包含了大量的审计记录,每一条记录具有 41 个属性,如协议类型、错误分段等^[12]。由于 KDD cup 数据集本身十分庞大,本课题从中提取了一部分数据形成实验数据集进行检测。

实验将 KDD cup 数据集中的数据作为图 2 所示的入侵检测方法的待检测的数据连接记录,对其进行分析、处理和聚类,并得出最终的入侵检测结果。实验输入多个不同的数据集进行了多次测试并统计了结果。

实验部分结果如图 4 所示,其中的每一个数值表示对应的所输入的 KDD cup 数据集中的每一条数据连接记录的编号。

其中,normal 表示正常的数据集。

probing 表示端口监视或扫描类的攻击。

dos 表示拒绝服务攻击。

U2R 表示未授权的本地超级用户特权访问。

R2L 表示来自远程主机的未授权访问。

此外,实验统计了局部收敛的情况,如表 1 所示。

表 1 局部收敛的次数占总实验次数比例对比

引入遗传算法之前	25.7%
引入遗传算法之后	2.8%

可以看出,在引入了遗传算法后,k-means 聚类算

法收敛于局部最优解的情况大大减少。可见,由于遗传算法的优良全局搜索性能可以使聚类算法尽可能避免陷入局部最优解,从而找到问题真正的全局最优解,并最终提高聚类的效率和质量,提高入侵检测的检测率、降低了误报率。

```

normal:
0.1.2.3.4.5.6.7.8.9.10.11.12.13.14.15.
30.31.32.33.34.35.36.37.38.39.40.41.42.
57.58.59.60.61.62.63.64.65.66.67.68.69.
3.84.85.86.87.88.89.90.91.92.93.94.95.
7.108.109.110.111.112.113.114.115.116.
7.128.129.130.131.132.133.134.135.136.

probing:
3035.3036.3037.3038.3039.3040.3041.
3051.3052.3053.3054.3055.3056.3057.
3067.3068.3069.3070.3071.3072.3073.
3083.3084.3085.3086.3087.3088.3089.
3099.3100.3101.3102.3103.3104.3105.

dos:
3508.3509.3510.3511.3512.3513.3514.
3524.3525.3526.3527.3528.3529.3530.
3540.3541.3542.3543.3544.3545.3546.
3556.3557.3558.3559.3560.3561.3562.
3572.3573.3574.3575.3576.3577.3578.

U2R:
2679.2680.2681.2682.2683.2684.2685.
2695.2696.2697.2698.2699.2700.2701.
2711.2712.2713.2714.2715.2716.2717.
2727.2728.2729.2730.2731.2732.2733.
2743.2744.2745.2746.2747.2748.2749.

R2L:
21596.21597.21598.21599.21600.21601.
609.21610.21611.21612.21613.21614.2.
2.21623.21624.21625.21626.21627.216.
21636.21637.21638.21639.21640.21641.
649.21650.21651.21652.21653.21654.2.

```

图 4 部分聚类结果

由于入侵检测系统的评估标准主要有检测率和误报率两个方面。实验从这两个方面对 GCAH 入侵检测方法的检测效果进行了评估。

对多次实验的结果进行平均后入侵检测的检测率和误报率结果如表 2 和表 3 所示。

表 2 采用 k-means 算法的测试结果

类型	检测率	误报率
Normal	75.3%	37.9%
Probe	67.3%	
DOS	62.7%	
U2R	23%	
R2L	35.4%	

表 3 采用 GCAH 方法后的测试结果

类型	检测率	误报率
Normal	97.6%	7.8%
Probe	96.2%	
DOS	92.3%	
U2R	80.7%	
R2L	87.9%	

综上所述,由于遗传算法的优良全局搜索性能可

以使聚类算法尽可能避免陷入局部最优解,从而找到问题真正的全局最优解,并最终提高聚类的效果和质量,提高入侵检测的检测率、降低了误报率。文中的入侵检测方法 GCAH 比单一的未经改进的 k-means 聚类算法的检测率更高,尤其是在对 U2R 和 R2L 的检测效果方面更为明显。说明文中的方法提高了聚类的效率和质量,提高了入侵检测的检测率、降低了误报率,可以作为一种有效的基于数据挖掘技术的人侵检测方法。

5 结束语

从数据挖掘角度研究网络入侵检测方法,采用聚类算法与遗传算法混合的入侵检测方法 GCAH 对数据进行分析 and 检测,以避免 k-means 聚类算法收敛于局部最优解的固有缺陷。最后进行了实验和测试,结果表明 GCAH 方法从很大程度上能避免局部收敛问题,提高检测率、降低误报率,提高效率,能达到预期的效果。

参考文献:

- [1] Zhang Yan, Ou Yangjia. The Design and Implementation of Host-based Intrusion Detection System [C]//Intelligent Information Technology and Security Information. [s. l.]: [s. n.], 2010: 595-598.
 - [2] Zhang Cuixiao, Zhang Guobing, Sun Shanshan. A Mixed Unsupervised Clustering-based Intrusion Detection Model [C]//3rd International Conference on Genetic and Evolutionary Computing. [s. l.]: [s. n.], 2009: 426-428.
 - [3] 鲜永菊. 入侵检测 [M]. 西安: 西安电子科技大学出版社, 2009.
 - [4] 赵 辉, 张 鹏. 网络异常的主动检测与特征分析 [J]. 计算机技术与发展, 2010, 20(8): 159-165.
 - [5] 彭铮良. 网络安全入侵检测系统 [J]. 计算机周刊, 2001 (24): 22-23.
 - [6] 王 鑫, 王洪国, 王 珺, 等. 数据挖掘中聚类方法比较研究 [J]. 计算机技术与发展, 2007, 17(10): 20-25.
 - [7] 王光宏, 蒋 平. 数据挖掘综述 [J]. 同济大学学报, 2004, 32(2): 246-252.
 - [8] 苏辉贵, 傅秀芬, 钟 洪, 等. 数据挖掘在入侵检测中的应用 [J]. 计算机技术与发展, 2006, 16(10): 143-148.
 - [9] 文小燕, 杜海若. 数据挖掘的发展和综述 [J]. 电脑知识与技术, 2007(18): 1486-1513.
 - [10] Han Jiawei, Kamber M. 数据挖掘概念与技术 (中译本) [M]. 北京: 机械工业出版社, 2006.
 - [11] Koza J R. Genetic Programming, on the Programming of Computers by Means of Natural Selection [M]. [s. l.]: MIT Press, 1992.
 - [12] KDD cup 99 data [EB/OL]. 1999. <http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html>.
-
- (上接第 156 页)
- ### 参考文献:
- [1] Pang Bo, Lee Lillian, Vaithyanathan S. Thumbs up? Sentiment Classification using Machine Learning Techniques [C]// In Proceedings of Conf. on EMNLP02. [s. l.]: [s. n.], 2002.
 - [2] Turney P. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews [C]// In Proc. of the Meeting of the Association for Computational Linguistics (ACL02). [s. l.]: [s. n.], 2002: 417-424.
 - [3] Dave K, Lawrence S, Pennock D. Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews [C]// In Proc. of the 12th Intl. World Wide Web Conference (WWW03). [s. l.]: [s. n.], 2003: 519-528.
 - [4] Pang Bo, Lee Lillian. A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts [C]// In Proceedings of the 42nd ACL. [s. l.]: [s. n.], 2004: 271-278.
 - [5] Kim S, Hovy E. Determining the Sentiment of Opinions [C]// In Proc. of the Intl. Conf. on Computational Linguistics (COLING04). [s. l.]: [s. n.], 2004.
 - [6] Liu B, Hu M. Opinion Observer: Analyzing and Comparing Opinions on the Web [C]// In Proc of the 14th Intl. Word Web Web Conf. (WWW05). [s. l.]: [s. n.], 2005: 342-351.
 - [7] Yi J, Nasukawa T, Bunescu R C, et al. Sentiment Analyzer: Extracting Sentiments about a Given Topic Using Natural Language Processing Techniques [C]// In Proc. of the IEEE Conf. on Data Mining (ICDM03). [s. l.]: [s. n.], 2003.
 - [8] Ku Lun-Wei, Liang YU-Ting, Chen Hsin-His. Opinion Extraction, Summarization and Tracking in News and Blog Corpora [C]// Proc. of AAAI 2006. [s. l.]: [s. n.], 2006: 280-288.
 - [9] Melville P, Gryc W, Larence R D. Sentiment Analysis of Blogs by Combining Lexical Knowledge with Text Classification [C]// Proceedings of KDD-09. [s. l.]: [s. n.], 2009: 1275-1283.
 - [10] Zhang W, Yu C T, Meng W. Opinion Retrieval from Blogs [C]// Proc. Of CIKM 2007. [s. l.]: [s. n.], 2007: 831-840.
 - [11] Mei Qiaozhu, Xu Ling, Wondra M, et al. Topic Sentiment Mixture: Modeling Facets and Opinions in Weblogs [C]// WWW2007. [s. l.]: [s. n.], 2007: 350-358.
 - [12] Harb A, Plantie M, Dray G. Web Opinion Mining: How to extract opinions from blogs [C]// CSTSC2008. [s. l.]: [s. n.], 2008: 320-326.
 - [13] Liu Bing. Web 数据挖掘 [M]. 俞 勇, 薛贵荣, 韩定一, 译. 北京: 清华大学出版社, 2009.