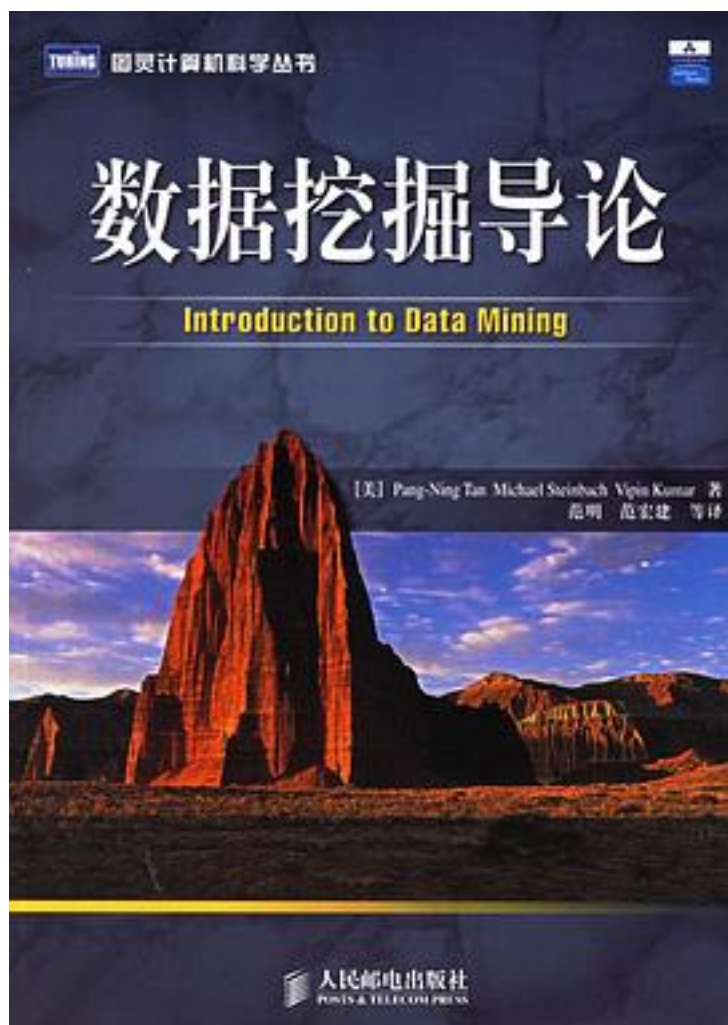


数据挖掘导论



[数据挖掘导论_下载链接1](#)

著者:(美)Pang-Ning Tan

出版者:机械工业出版社

出版时间:2010-9

装帧:

isbn:9787111316701

本书全面介绍了数据挖掘的理论和方法，着重介绍如何用数据挖掘知识解决各种实际问

题，涉及学科领域众多，适用面广。

书中涵盖5个主题：数据、分类、关联分析、聚类和异常检测。除异常检测外，每个主题都包含两章：前面一章讲述基本概念、代表性算法和评估技术，后面一章较深入地讨论高级概念和算法。目的是使读者在透彻地理解数据挖掘基础的同时，还能了解更多重要的高级主题。

本书特色

- 包含大量的图表、综合示例和丰富的习题。
- 不需要数据库背景，只需要很少的统计学或数学背景知识。
- 网上配套教辅资源丰富，包括ppt、习题解答、数据集等。

作者介绍:

Pang-Ning

Tan现为密歇根州立大学计算机与工程系助理教授，主要教授数据挖掘、数据库系统等课程。他的研究主要关注于为广泛的应用(包括医学信息学、地球科学、社会网络、Web挖掘和计算机安全)开发适用的数据挖掘算法。

Michael

Steinbach拥有明尼苏达大学数学学士学位、统计学硕士学位和计算机科学博士学位，现为明尼苏达大学双城分校计算机科学与工程系助理研究员。

Vipin Kumar现为明尼苏达大学计算机科学与工程系主任和William Norris教授。1988年至2005年，他曾担任美国陆军高性能计算研究中心主任。

目录: preface v

1 introduction 1

1.1 what is data mining 2

1.2 motivating challenges 4

1.3 the origins of data mining 6

1.4 data mining tasks 7

1.5 scope and organization of the book 11

1.6 bibliographic notes 13

1.7 exercises 16

2 data 19

2.1 types of data 22

2.1.1 attributes and measurement 23

2.1.2 types of data sets 29

2.2 data quality 36

2.2.1 measurement and data collection issues 37

2.2.2 issues related to applications 43

2.3 data preprocessing 44

2.3.1 aggregation 45

2.3.2 sampling 47

2.3.3 dimensionality reduction 50

2.3.4 feature subset selection 52

2.3.5 feature creation 55

2.3.6 discretization and binarization 57

2.3.7 variable transformation	63
2.4 measures of similarity and dissimilarity	65
2.4.1 basics	66
2.4.2 similarity and dissimilarity between simple attributes	67
2.4.3 dissimilarities between data objects	69
2.4.4 similarities between data objects	72
2.4.5 examples of proximity measures	73
2.4.6 issues in proximity calculation	80
2.4.7 selecting the right proximity measure	83
2.5 bibliographic notes	84
2.6 exercises	88
3 exploring data	97
3.1 the iris data set	98
3.2 summary statistics	98
3.2.1 frequencies and the mode	99
3.2.2 percentiles	100
3.2.3 measures of location: mean and median	101
3.2.4 measures of spread: range and variance	102
3.2.5 multivariate summary statistics	104
3.2.6 other ways to summarize the data	105
3.3 visualization	105
3.3.1 motivations for visualization	105
3.3.2 general concepts	106
3.3.3 techniques	110
3.3.4 visualizing higher-dimensional data	124
3.3.5 do's and don'ts	130
3.4 olap and multidimensional data analysis	131
3.4.1 representing iris data as a multidimensional array	131
3.4.2 multidimensional data: the general case	133
3.4.3 analyzing multidimensional data	135
3.4.4 final comments on multidimensional data analysis	139
3.5 bibliographic notes	139
3.6 exercises	141
4 classification:	
basic concepts, decision trees, and model evaluation	145
4.1 preliminaries	146
4.2 general approach to solving a classification problem	148
4.3 decision tree induction	150
4.3.1 how a decision tree works	150
4.3.2 how to build a decision tree	151
4.3.3 methods for expressing attribute test conditions	155
4.3.4 measures for selecting the best split	158
4.3.5 algorithm for decision tree induction	164
4.3.6 an example: web robot detection	166
contents	xi
4.3.7 characteristics of decision tree induction	168
4.4 model overfitting	172
4.4.1 overfitting due to presence of noise	175
4.4.2 overfitting due to lack of representative samples	177
4.4.3 overfitting and the multiple comparison procedure	178
4.4.4 estimation of generalization errors	179
4.4.5 handling overfitting in decision tree induction	184
4.5 evaluating the performance of a classifier	186
4.5.1 holdout method	186

4.5.2 random subsampling	187
4.5.3 cross-validation	187
4.5.4 bootstrap	188
4.6 methods for comparing classifiers	188
4.6.1 estimating a confidence interval for accuracy	189
4.6.2 comparing the performance of two models	191
4.6.3 comparing the performance of two classifiers	192
4.7 bibliographic notes	193
4.8 exercises	198
5 classification: alternative techniques	207
5.1 rule-based classifier	207
5.1.1 how a rule-based classifier works	209
5.1.2 rule-ordering schemes	211
5.1.3 how to build a rule-based classifier	212
5.1.4 direct methods for rule extraction	213
5.1.5 indirect methods for rule extraction	221
5.1.6 characteristics of rule-based classifiers	223
5.2 nearest-neighbor classifiers	223
5.2.1 algorithm	225
5.2.2 characteristics of nearest-neighbor classifiers	226
5.3 bayesian classifiers	227
5.3.1 bayes theorem	228
5.3.2 using the bayes theorem for classification	229
5.3.3 naïve bayes classifier	231
5.3.4 bayes error rate	238
5.3.5 bayesian belief networks	240
5.4 artificial neural network (ann)	246
5.4.1 perceptron	247
5.4.2 multilayer artificial neural network	251
5.4.3 characteristics of ann	255
xii contents	
5.5 support vector machine (svm)	256
5.5.1 maximum margin hyperplanes	256
5.5.2 linear svm: separable case	259
5.5.3 linear svm: nonseparable case	266
5.5.4 nonlinear svm	270
5.5.5 characteristics of svm	276
5.6 ensemble methods	276
5.6.1 rationale for ensemble method	277
5.6.2 methods for constructing an ensemble classifier	278
5.6.3 bias-variance decomposition	281
5.6.4 bagging	283
5.6.5 boosting	285
5.6.6 random forests	290
5.6.7 empirical comparison among ensemble methods	294
5.7 class imbalance problem	294
5.7.1 alternative metrics	295
5.7.2 the receiver operating characteristic curve	298
5.7.3 cost-sensitive learning	302
5.7.4 sampling-based approaches	305
5.8 multiclass problem	306
5.9 bibliographic notes	309
5.10 exercises	315
6 association analysis: basic concepts and algorithms	327

6.1	problem definition	328
6.2	frequent itemset generation	332
6.2.1	the apriori principle	333
6.2.2	frequent itemset generation in the apriori algorithm	335
6.2.3	candidate generation and pruning	338
6.2.4	support counting	342
6.2.5	computational complexity	345
6.3	rule generation	349
6.3.1	confidence-based pruning	350
6.3.2	rule generation in apriori algorithm	350
6.3.3	an example: congressional voting records	352
6.4	compact representation of frequent itemsets	353
6.4.1	maximal frequent itemsets	354
6.4.2	closed frequent itemsets	355
6.5	alternative methods for generating frequent itemsets	359
6.6	fp-growth algorithm	363
contents xiii		
6.6.1	fp-tree representation	363
6.6.2	frequent itemset generation in fp-growth algorithm	366
6.7	evaluation of association patterns	370
6.7.1	objective measures of interestingness	371
6.7.2	measures beyond pairs of binary variables	382
6.7.3	simpson's paradox	384
6.8	effect of skewed support distribution	386
6.9	bibliographic notes	390
6.10	exercises	404
7	association analysis: advanced concepts	415
7.1	handling categorical attributes	415
7.2	handling continuous attributes	418
7.2.1	discretization-based methods	418
7.2.2	statistics-based methods	422
7.2.3	non-discretization methods	424
7.3	handling a concept hierarchy	426
7.4	sequential patterns	429
7.4.1	problem formulation	429
7.4.2	sequential pattern discovery	431
7.4.3	timing constraints	436
7.4.4	alternative counting schemes	439
7.5	subgraph patterns	442
7.5.1	graphs and subgraphs	443
7.5.2	frequent subgraph mining	444
7.5.3	apriori-like method	447
7.5.4	candidate generation	448
7.5.5	candidate pruning	453
7.5.6	support counting	457
7.6	infrequent patterns	457
7.6.1	negative patterns	458
7.6.2	negatively correlated patterns	458
7.6.3	comparisons among infrequent patterns, negative patterns, and negatively correlated patterns	460
7.6.4	techniques for mining interesting infrequent patterns	461
7.6.5	techniques based on mining negative patterns	463
7.6.6	techniques based on support expectation	465
7.7	bibliographic notes	469

7.8 exercises	473
xiv contents	
8 cluster analysis: basic concepts and algorithms	487
8.1 overview	490
8.1.1 what is cluster analysis	490
8.1.2 different types of clusterings	491
8.1.3 different types of clusters	493
8.2 k-means	496
8.2.1 the basic k-means algorithm	497
8.2.2 k-means: additional issues	506
8.2.3 bisecting k-means	508
8.2.4 k-means and different types of clusters	510
8.2.5 strengths and weaknesses	510
8.2.6 k-means as an optimization problem	513
8.3 agglomerative hierarchical clustering	515
8.3.1 basic agglomerative hierarchical clustering algorithm	516
8.3.2 specific techniques	518
8.3.3 the lance-williams formula for cluster proximity	524
8.3.4 key issues in hierarchical clustering	524
8.3.5 strengths and weaknesses	526
8.4 dbscan	526
8.4.1 traditional density: center-based approach	527
8.4.2 the dbscan algorithm	528
8.4.3 strengths and weaknesses	530
8.5 cluster evaluation	532
8.5.1 overview	533
8.5.2 unsupervised cluster evaluation using cohesion and separation	536
8.5.3 unsupervised cluster evaluation using the proximity matrix	542
8.5.4 unsupervised evaluation of hierarchical clustering	544
8.5.5 determining the correct number of clusters	546
8.5.6 clustering tendency	547
8.5.7 supervised measures of cluster validity	548
8.5.8 assessing the significance of cluster validity measures	553
8.6 bibliographic notes	555
8.7 exercises	559
9 cluster analysis: additional issues and algorithms	569
9.1 characteristics of data, clusters, and clustering algorithms	570
9.1.1 example: comparing k-means and dbscan	570
9.1.2 data characteristics	571
contents	xv
9.1.3 cluster characteristics	573
9.1.4 general characteristics of clustering algorithms	575
9.2 prototype-based clustering	577
9.2.1 fuzzy clustering	577
9.2.2 clustering using mixture models	583
9.2.3 self-organizing maps (som)	594
9.3 density-based clustering	600
9.3.1 grid-based clustering	601
9.3.2 subspace clustering	604
9.3.3 denclue: a kernel-based scheme for density-based clustering	608
9.4 graph-based clustering	612

- 9.4.1 sparsification 613
- 9.4.2 minimum spanning tree (mst) clustering 614
- 9.4.3 opossum: optimal partitioning of sparse similarities using metis 616
- 9.4.4 chameleon: hierarchical clustering with dynamic modeling 616
- 9.4.5 shared nearest neighbor similarity 622
- 9.4.6 the jarvis-patrick clustering algorithm 625
- 9.4.7 snn density 627
- 9.4.8 snn density-based clustering 629
- 9.5 scalable clustering algorithms 630
 - 9.5.1 scalability: general issues and approaches 630
 - 9.5.2 birch 633
 - 9.5.3 cure 635
- 9.6 which clustering algorithm 639
- 9.7 bibliographic notes 643
- 9.8 exercises 647
- 10 anomaly detection 651
 - 10.1 preliminaries 653
 - 10.1.1 causes of anomalies 653
 - 10.1.2 approaches to anomaly detection 654
 - 10.1.3 the use of class labels 655
 - 10.1.4 issues 656
 - 10.2 statistical approaches 658
 - 10.2.1 detecting outliers in a univariate normal distribution 659
 - 10.2.2 outliers in a multivariate normal distribution 661
 - 10.2.3 a mixture model approach for anomaly detection 662
 - xvi contents
 - 10.2.4 strengths and weaknesses 665
 - 10.3 proximity-based outlier detection 666
 - 10.3.1 strengths and weaknesses 666
 - 10.4 density-based outlier detection 668
 - 10.4.1 detection of outliers using relative density 669
 - 10.4.2 strengths and weaknesses 670
 - 10.5 clustering-based techniques 671
 - 10.5.1 assessing the extent to which an object belongs to a cluster 672
 - 10.5.2 impact of outliers on the initial clustering 674
 - 10.5.3 the number of clusters to use 674
 - 10.5.4 strengths and weaknesses 674
 - 10.6 bibliographic notes 675
 - 10.7 exercises 680
- appendix a linear algebra 685
 - a.1 vectors 685
 - a.1.1 definition 685
 - a.1.2 vector addition and multiplication by a scalar 685
 - a.1.3 vector spaces 687
 - a.1.4 the dot product, orthogonality, and orthogonal projections 688
 - a.1.5 vectors and data analysis 690
 - a.2 matrices 691
 - a.2.1 matrices: definitions 691
 - a.2.2 matrices: addition and multiplication by a scalar 692
 - a.2.3 matrices: multiplication 693

a.2.4 linear transformations and inverse matrices 695
a.2.5 eigenvalue and singular value decomposition 697
a.2.6 matrices and data analysis 699
a.3 bibliographic notes 700
appendix b dimensionality reduction 701
b.1 pca and svd 701
b.1.1 principal components analysis (pca) 701
b.1.2 svd 706
b.2 other dimensionality reduction techniques 708
b.2.1 factor analysis 708
b.2.2 locally linear embedding (lle) 710
b.2.3 multidimensional scaling, fastmap, and isomap 712
contents xvii
b.2.4 common issues 715
b.3 bibliographic notes 716
appendix c probability and statistics 719
c.1 probability 719
c.1.1 expected values 722
c.2 statistics 723
c.2.1 point estimation 724
c.2.2 central limit theorem 724
c.2.3 interval estimation 725
c.3 hypothesis testing 726
appendix d regression 729
d.1 preliminaries 729
d.2 simple linear regression 730
d.2.1 least square method 731
d.2.2 analyzing regression errors 733
d.2.3 analyzing goodness of fit 735
d.3 multivariate linear regression 736
d.4 alternative least-square regression methods 737
appendix e optimization 739
e.1 unconstrained optimization 739
e.1.1 numerical methods 742
e.2 constrained optimization 746
e.2.1 equality constraints 746
e.2.2 inequality constraints 747
author index 750
subject index 758
copyright permissions 769
xviii contents
• • • • • ([收起](#))

[数据挖掘导论 下载链接1](#)

标签

数据挖掘

机器学习

算法

Data-Mining

计算机科学

计算机

数据研究

Mining

评论

很后悔在学校时没有多读英文的原版书，以致于当时上完课之后对很多知识的认知都是停留在知其然而不知其所以然的阶段。英文原版书真的是好太多，来龙去脉讲的很清楚，不仅是知其然，更重要的是能提升对问题的研究兴趣和思考能力。当然，这本书只是 general introduction，后面还得不断地深挖才行。

春节做个笔记吧。感觉SVM，神经网络和BBN都讲的有些浅尝即止了

浅尝辄止,可以拿来当科普...

粗粗的浏览了一遍，把各种方法的逻辑熟悉了一下，下学期信科好像有门这个课，去好好学一下。真个挺难的。不过很有用，要好好学。并且要配合这一年的基础数学学习。

A solid textbook. And perhaps I should build some simple projects meanwhile reading

through it. Acturally I didn't, so when I read the latter half, it became boring somehow. Finding and Solving a bit problems immediately sounds good.

原版，非常详细

写的很详细

各方面都很不错的书

入门好书，写论文的时候参考了

2015年初读完。很赞，第一本细读的ai related book. 英文纸质版。基本通读过一遍，个别章节跳过了。

经典~

查的时候经常找不到…

Go Data Mining.

英文原版，通读后，数据挖掘的理论基础

不错，基础又相对系统 另: 中文版太lj，建议直接英文版

好教材

[数据挖掘导论_下载链接1](#)

书评

我的习惯就是在蹲坑的时候读一些艰涩高深的科学读物，这样有助于我在排泄的时候大脑保持高度的兴奋状态，不至于被熏晕或者不至于被引人入胜的小说情节所陶醉最后导致肛痿……但是，这本书另我惊诧了……
第一他不艰涩，是我读到过的关于统计、关于数据、关于计算的最科普的读...

该书特点：以实例为重，给出了常用算法的伪代码，和《模式识别》、《模式分类》等专著比起来，该书略去了各个定理的证明部分，并通过大量枚举具体的分类实例，来简要说明算法的流程和意义。
根据个人的体验，觉得这本书作为第一本数据挖掘的入门读物是再恰当不过的了。...

这本书介绍的比较全面，某些内容在一般的书中是很少介绍的，内容浅显易懂。本人开始看中文版的，觉的中文版的写的不错，后来又看英文版的，就发现中文版的差太多了，推荐英文版的

我是拿这本书当作课程书的，这本书基本上涵盖了数据挖掘的许多经典算法，分类，聚类，关联规则。比较适合对数据挖掘感兴趣的人，这本书看完之后基本上就可以进行对数据的分析，挖掘了。然而这仅仅是一门入门书，对于理论部分并没有做过多的解释。如果想进一步的了解理论知识，...

作为数据挖掘导论，这本书基本上已经做到了。书中介绍了很多数据挖掘方面相关的概念和方法，对于入门来讲是很友好的。因为刚刚看完机器学习的书，所以前半部分基本不需要看了。后面的关联分析和聚类方法还是可以一看的。虽然这本书没有实际操作的

Chapter2 和 Chapter3

一大堆废话，基本都是初中高中教的！！！好像跳过这些章节！！！ Chapter2 和 Chapter3 一大堆废话，基本都是初中高中教的！！！好像跳过这些章节！！！！

Chapter2 和 Chapter3

一大堆废话，基本都是初中高中教的！！！好像跳过这些章节！！！！

它是我关于数据挖掘这一方向的入门书。

书中讲了很多基础的数据挖掘算法，读完以后可以对这些算法的基本思想有个了解。书中的例子也很详尽，还是不错的。

但是研究生期间是指望发论文的，这些算法从学术上来说，只能算基础入门了。至于它们在实际工业应...

这本书写得逻辑性比较强，全面，而且我觉得涉及的东西也比较底层，让我们了解一些算法的基本型原理是非常重要的。如果，网上的机器学习相关文章看不懂的话，可以从这本书入手。中文版的只看过一点点，感觉完全没逻辑性，完全没感觉。翻译出来完全就变味了，毕竟是语言习惯上的...

给出了DataMining的一般性解决思路，全面易懂，很适合给初学者扫盲。加之与原版大概400+RMB比较起来，不禁觉得还是祖国好哇。。。PS:据说巴基斯坦卖得更便宜。。。

看我截图吧 <http://weibo.com/1677386655/zu8O4ci9O> therefore, if we compute the k-dist for all the data points for some k, sort them in increasing order, and then plot the sorted values, we expect to see a sharp change at the value of k-dist that correspon...

[数据挖掘导论_下载链接1](#)