中国计算机学会学术著作丛书
——知识科学系列 10

机器学习及其应用2013

张长水 杨 强 主编

清华大学出版社

中国计算机学会学术著作丛书
——知识科学系列 10

机器学习及其应用2013

张长水 杨 强 主编

师范学院图书馆
藏书

清华大学出版社
北京

内 容 简 介

机器学习是计算机科学和人工智能中非常重要的一个研究领域。近年来，机器学习不仅在计算机科学的众多领域中大显身手，还成为一些交叉学科的重要支持技术。本书邀请国内外相关领域的专家撰文，以综述的形式分别介绍机器学习不同分支及相关领域的研究进展。全书共分 8 章，内容分别涉及稀疏话题表示学习、基于向量场的流形学习和排序、秩极小化、实值多变量维数约简等技术，知识挖掘与用户建模、异质人脸图像合成等应用，以及对多视图在利用未标记数据学习中的效用、面向高维多视图数据的广义相关分析的探讨。

本书可供高校、科研院所计算机、自动化及相关专业的师生、科技工作者和相关企业的工程技术人员阅读参考。

版权所有，侵权必究。侵权举报电话：010-62782989　13701121993

图书在版编目(CIP)数据

机器学习及其应用. 2013/张长水，杨强主编. —北京：清华大学出版社，2013
(中国计算机学会学术著作丛书. 知识科学系列)
ISBN 978-7-302-33619-8

Ⅰ. ①机… Ⅱ. ①张… ②杨… Ⅲ. ①机器学习 Ⅳ. ①TP181

中国版本图书馆 CIP 数据核字(2013)第 203926 号

责任编辑：薛　慧
封面设计：常雪影
责任校对：刘玉霞
责任印制：宋　林

出版发行：清华大学出版社
网　　址：http://www.tup.com.cn，http://www.wqbook.com
地　　址：北京清华大学学研大厦 A 座　　邮　　编：100084
社 总 机：010-62770175　　邮　　购：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质 量 反 馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
印 刷 者：北京富博印刷有限公司
装 订 者：北京市密云县京文制本装订厂
经　　销：全国新华书店
开　　本：185mm×230mm　**印　张**：13.5　**插　页**：1　　**字　　数**：261 千字
版　　次：2013 年 10 月第 1 版　　**印　　次**：2013 年 10 月第1次印刷
印　　数：1～2200
定　　价：43.00 元

产品编号：054999-01

前　言

机器学习致力于“利用经验来改善系统自身的性能”。在计算机系统中，“经验”通常是以数据的形式存在的，要利用经验就不可避免地要对数据进行分析，因此，机器学习已逐渐成为计算机数据分析技术的源泉之一。随着人类收集和存储数据能力的不断增长以及计算机运算能力的飞速发展，利用计算机来分析数据的要求越来越广泛，越来越迫切，从而使得机器学习的重要性越来越显著。机器学习不仅是人工智能的核心研究领域之一，目前还成为计算机科学中最活跃、最受关注的领域之一。

2002 年，陆汝钤院士在复旦大学智能信息处理实验室发起组织了“智能信息处理系列研讨会”，并将“机器学习及其应用”列为当年支持的研讨会之一。2002 年 11 月，研讨会成功举行，并确定了会议不征文、不收费、报告人由组织者邀请，以及“学术至上，其他从简”的办会宗旨。2004 年 11 月，在复旦大学举行了第二届“机器学习及其应用”研讨会，两天半的会议一直有 100 余人旁听。2005 年起，研讨会由南京大学软件新技术国家重点实验室举办。2005 年 11 月举办的第三届研讨会吸引了来自全国近 10 个省市的 250 余人旁听；2006 年 11 月、2007 年 11 月分别由南京航空航天大学信息科学与技术学院、南京师范大学数学与计算机学院协办第四届和第五届研讨会，两次均吸引了来自全国 10 余个省市的约 300 人旁听；2008 年 11 月举行的第六届研讨会，适逢南京大学计算机学科建立 50 周年，吸引了来自全国 10 余个省市的 380 余人旁听；此后在 2009 年 11 月和 2010 年 11 月在南京大学分别举行了第七、八届研讨会，均有约 400 人旁听。2011 年 11 月和 2012 年 11 月由清华大学自动化系、智能科学与系统国家重点实验室、清华大学信息科学与技术国家实验室(筹)举办第九届和第十届研讨会，这两次会议均有 500 多人旁听。

清华大学出版社对推介信息科学技术领域的研究进展一直抱有极大的热情。早在“第二届机器学习及其应用研讨会”举行时清华大学出版社就参与其中，并为该研讨会专门出版了文集，即 2006 年的《机器学习及其应用》。2005 年第三届研讨会期间，出版社和与会专家商定，以后每两届研讨会的部分内容将汇编结集，以《机器学习及其应用＋出版

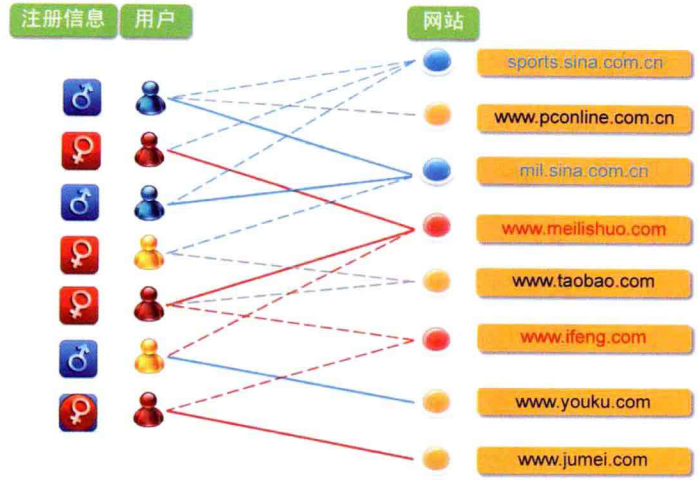
年》的形式冠名。第三至八届研讨会的部分内容已在《机器学习及其应用 2007》、《机器学习及其应用 2009》以及《机器学习及其应用 2011》中出版面世。

本书是清华大学出版社邀请第九届和第十届“机器学习及其应用研讨会”的部分专家将报告内容总结而形成的文集。书中每篇文章讨论一个方面的问题，以综述形式介绍这个方面的研究工作，包括自己的研究工作。本书收录的 8 篇文章，每一位作者都投入了大量的时间和精力，深入浅出地介绍了一个领域的来龙去脉，并讨论其发展趋势。本书的出版得到了陆汝钤院士、王珏老师的支持和指导，并得到清华大学出版社计算机专著出版基金的资助，在此一并表示衷心的感谢。

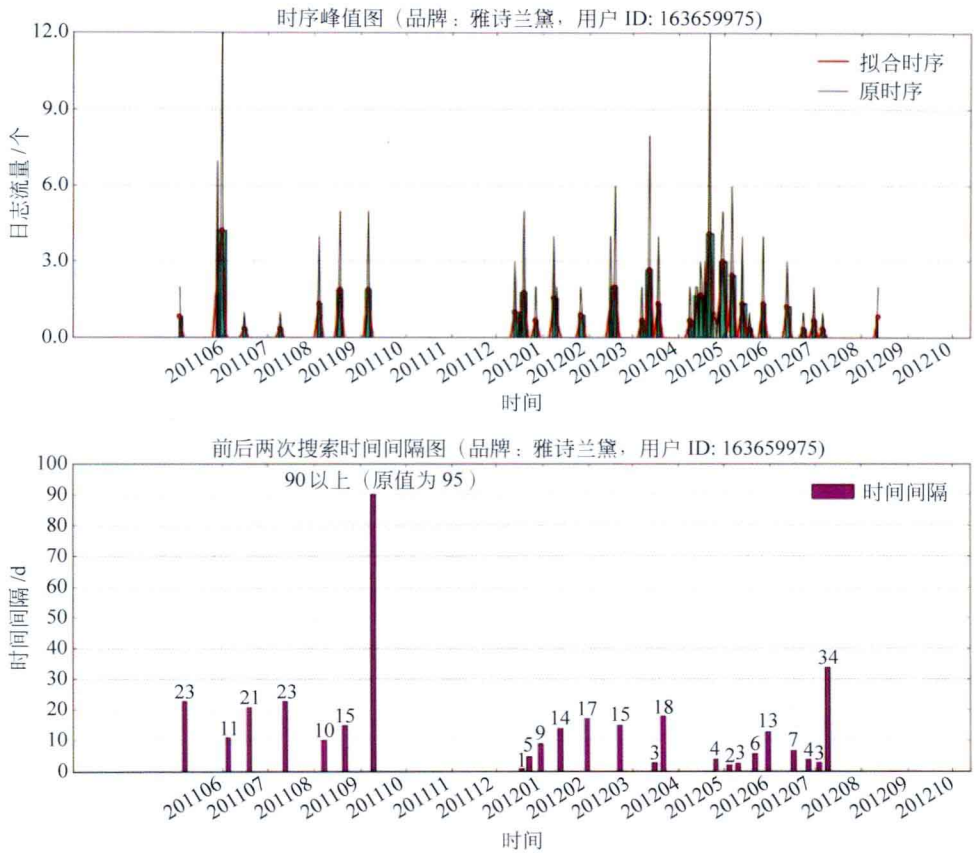
编　者

2013 年 8 月

此为试读，需要完整PDF请访问：www.ertongbook.com

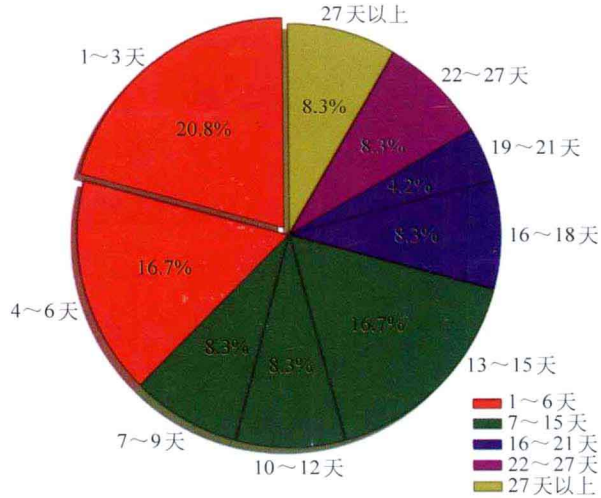


彩插 1　基于 Co-training 技术的用户属性模型



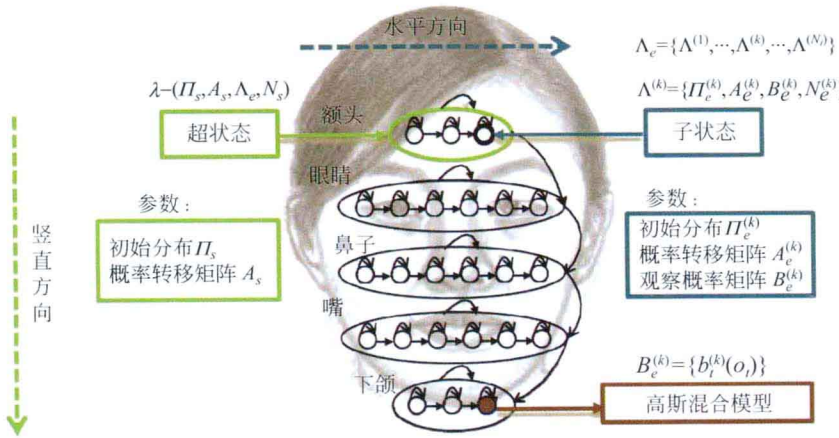
彩插 2　个体用户搜索时间间隔分布图

该图由上下两个子图组成，上方子图为其对应的时序峰值检测图，其中绿色直方图描述时序峰值的大小和覆盖范围，下方子图为根据该用户的搜索行为时间模式检测出来其对应前后 2 个集中关注的时间间隔分布图，其中紫色直方图描述峰值之间的时间间隔的起始位置和间隔的时间长度大小

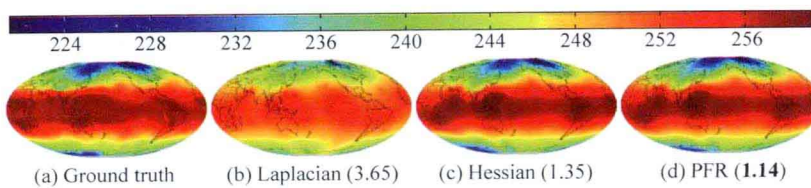


彩插 3　个体用户未来消费概率预判图

五种颜色代表五种间隔范围：红色代表搜索周期介于 1～6 天、绿色代表搜索周期介于 7～15 天、蓝色代表搜索周期介于 16～21 天、紫色代表搜索周期介于 22～27 天、黄色代表搜索周期大于 27 天



彩插 4　嵌入式隐马尔科夫模型示意图



彩插 5　全球温度预测（详细说明见本书第 141 页）

目　录

此为试读,需要完整PDF请访问: www.ertongbook.com

Learning Sparse Topical Representations

Jun Zhu[†] Aonan Zhang[†] Eric P. Xing[‡]
[†]Dept. of CS & T, TNList Lab, State Key Lab of ITS, Tsinghua University, Beijing 100084, China
{dcszj, zan12}@mail. tsinghua. edu. cn
[‡]School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA
epxing@cs. cmu. edu

1 Introduction

Learning a representation that captures the latent semantics of a large collection of data is an important problem in many scientific and engineering applications. Probabilistic topic models such as LDA (latent Dirichlet allocation) [Blet *et al.* 2003] posits that each document is an admixture of latent topics where each topic is a unigram distribution over the terms in a vocabulary. The document-specific admixture proportion vector can be regarded as a representation of the document in the latent topic space, which can be used for classification [Zhu *et al.* 2009], retrieval [Hofmann 1999] or visualizing the otherwise unstructured collection; and the inferred word-level topic assignment distributions can be useful for word sense induction [Brody and Lapata 2009] or disambiguation [Boyd-Graber *et al.* 2007].

However, such a probabilistic topic model is largely limited in two aspects. First, it lacks a mechanism to explicitly control the sparsity of the inferred representations. Sparsity of the representations in a semantic space is a desirable property in text modeling[Shashanka *et al.* 2007; Wang and Blei 2009] and human vision [Olshausen and

Field 1996]. For example, very often it makes intuitive sense to assume that each document or each word has a few salient topical meanings or senses [Shashanka *et al.* 2007; Wang and Blei 2009], rather than letting every topic or sense make a non-zero contribution; this is especially important in practice for large scale text mining endeavors such as those undertaken in Google or Yahoo, where it is not uncommon to learn hundreds if not thousands of topics for hundreds of millions of documents—without an explicit sparcification procedure, it would be extremely challenging, if not impossible, to nail down the semantic meanings of a document or word. Second, the probabilistic nature of such topic models could make it computationally difficult to incorporate supervised side information [Wang *et al.* 2009] or a rich set of features [Zhu and Xing 2010]. This is because each component in such a probabilistic model needs to be a normalized distribution, in which the normalization factor or log-partition function could make the inference extremely hard.

To achieve sparsity in a probabilistic topic model is non-trivial. Existing attempts, such as imposing posterior regularization (e. g. , using entropic priors[Shashanka *et al.* 2007] or moment constraints [Ganchev *et al.* 2010]), introducing auxiliary variables [Wang and Blei 2009], or using a sparse exponential prior in LDA [Yang *et al.* 2010], can in principle introduce a bias toward a posterior distribution that is concentrated on a small number of components (e. g. , topics). However, due to the smoothness of the regularizer (e. g. , entropic regularizer) or uncertainty of auxiliary variables, such methods often do not yield truly sparse results in practice. Moreover, the aforementioned methods aim either at achieving sparse document-level representations [Shashanka *et al.* 2007; Yang *et al.* 2010] or sparse topic vectors [Wang and Blei 2009]. To the best of our knowledge, no systematical study exists on discovering sparse word-level representations. For the second limitation, the non-probabilistic latent variable/factor models, such as non-negative matrix factorization (NMF) [Lee and Seung 1999] and sparse coding (SPC) methods, provide inspiring ideas to relax the strict normalization condition in probabilistic models.

As we have stated, the reason for the second limitation is that probabilistic models require to define normalized distributional components. Similarly, a technical reason for the difficulty in achieving sparsity in a probabilistic topic model is also that the admixing proportions or topics take the form as a normalized vector that defines a distribution.

Therefore, it is unhelpful to directly use a sparsity inducing ℓ_1-regularizer as in Lasso [Tibshirani 1996; Meinshausen and Yu, 2009]. In contrast, the non-probabilistic sparse coding [Olshausen and Field 1996] provides an elegant framework to achieve sparsity on the usually un-normalized code vector or dictionary (i. e., a basis set) by using the theoretically sound ℓ_1-regularizer or other composite regularizers [Kim and Xing 2010; Jenatton *et al.* 2010; Jacob *et al.* 2009; Bengio *et al.* 2009]. Although much work has been done on learning a structured dictionary [Jenatton *et al.* 2010; Bengio *et al.* 2009], existing sparse coding methods typically discover flat representations, such as the single-layer sparse codes of small image patches or word terms [Jenatton *et al.* 2010; Bengio *et al.* 2009]. In order to achieve a representation of an entire image or document from the sparse codes of its components, a post-processing such as average or max pooling [Yang *et al.* 2009] is needed. This two-step procedure can be rather sub-optimal because it lacks a channel to provide direct correlations between individual component representations [Hyvärinen and Hoyer 2001], or to leverage the possibly available high-level weak supervision (e. g., document categories) to discover predictive representations [Zhu *et al.* 2009] or learn a supervised dictionary [Mairal *et al.* 2008].

To address the above limitations, we present sparse topical coding (STC), a novel statistical method for learning sparse hierarchical representations of input samples, such as text documents. In STC, each noisy individual input feature (e. g., a word count) is reconstructed from a sparse linear combination of a set of bases, and the representation of an entire document is derived via an aggregation strategy (e. g., averaging or truncated averaging) from the sparse codes of all its individual word features. By using a log-loss under the broad exponential family of distributions, STC can model both discrete and continuous data. When applied to text, we use the log-Poisson loss to model discrete word counts and learn the bases that are unigram distributions over the terms in a vocabulary, also known as topics. We present an efficient coordinate descent algorithm to solve the hierarchical sparse coding problem, and the dictionary learning is efficiently done with projected gradient descent. Our algorithm provides a systematic (both algorithmic and empirical) comparison between STC and probabilistic LDA models [Blet *et al.* 2003].

In addition, we also describe a supervised STC (MedSTC) to show how to incorporate supervising side-information when it is available into the STC to discover

more predictive representations and learn a more accurate document classifier. Finally, we provide some empirical studies on text modeling and classification. Our results show that STC can learn meaningful topical bases, infer sparse topical representations of documents, and identify sparse topical senses of words which would be useful for word sense induction [Pantel and Lin 2002; Brody and Lapata 2009] or disambiguation [Boyd-Graber *et al.* 2007]. We report that both the unsupervised STC and supervised MedSTC outperform several competing methods on document classification and are significantly more efficient (an order of magnitude speed up) on training and testing.

This chapter is structured as follows. Section 2 introduces related work. Section 3 presents STC and an efficient coordinate descent algorithm. Section 4 describes a collapsed version of STC and MedSTC. Section 5 presents empirical studies, and Section 6 concludes with future directions discussed.

2 Related Work

Sparse coding is a powerful technique that can learn a generic dictionary from an unlabeled corpus. The learned dictionary can be further used to encode a data sample and find a new representation, which is useful for visualization, clustering, classification, or self-taught learning [Raina *et al.* 2007]. However, by treating the inputs as independent samples and using the flat ℓ_1-norm regularizer, the standard sparse coding has limitations because of its incapacity to learn structured dictionary and structured representations of the input samples. Much work has been done focusing on addressing the first problem to learn structured dictionary, such as [Jenatton *et al.* 2010] by using a structured sparsity regularizer (e. g. , group-wise Lasso [Jacob *et al.* 2009] or tree-guided Lasso [Kim and Xing 2010]), [Varshney *et al.* 2008] by designing a structure among dictionary elements, or [Jost *et al.* 2006] by using a clustering algorithm to construct a tree structure. However, much less work has been done on learning structured sparse representations of input samples. Sparse topical coding is a hierarchical sparse coding technique, and it has close relationships with latent Dirichlet allocation (LDA) [Blet *et al.* 2003] and non-negative matrix factorization (NMF) [Lee and Seung 1999], as detailed below.

2.1 Probabilistic LDA

STC is a hierarchical sparse coding method that shares the similar goal as the probabilistic LDA [Blet *et al.* 2003] for inferring latent representations of text documents. Before formally introducing STC, we discuss potential drawbacks of LDA on the model aspect.

First, LDA does not have an explicit definition of word code. In LDA, a document is represented as a *sequence* of words $\tilde{\boldsymbol{w}} = (w_1, w_2, \cdots, w_M)$, where M denotes document length and $\boldsymbol{w}_m$ is an N-dimensional indicator vector, that is, $w_{mn} = 1$ if word n appears in the mth position of the document; otherwise 0. LDA associates each position m with a topic assignment indicator variable Z_m and assumes that the topics of all the words in a document are sampled from the same document-level topic mixing proportion, which will be denoted by $\tilde{\boldsymbol{\theta}}$. By assuming a Dirichlet prior over the topic mixing proportion $\tilde{\boldsymbol{\theta}}$, LDA defines a joint distribution for a document

$$p(\tilde{\boldsymbol{\theta}}, \boldsymbol{z}, \tilde{\boldsymbol{w}} \mid \alpha, \beta) = p(\tilde{\boldsymbol{\theta}} \mid \alpha) \prod_{m=1}^{M} p(z_m \mid \tilde{\boldsymbol{\theta}}) p(\boldsymbol{w}_m \mid z_m, \beta) \tag{1}$$

where both the topic assignment model $p(z_m \mid \tilde{\boldsymbol{\theta}})$ and the word generating model $p(\boldsymbol{w}_m \mid z_m, \beta)$ are normalized multinomial *distributions* and α are Dirichlet parameters. For comparison, an equivalence to word code can be defined as the *empirical* word-topic assignment distribution $\tilde{p}(z(n)=k) \propto \sum_m w_{mn} q(z_{mk}=1 \mid \tilde{\boldsymbol{w}})$, where $z(n)$ is the topic of word n. The distribution $\tilde{p}(z(n))$ can be regarded as a representation of word n in the topic space, and it can be inferred using sampling [Brody and Lapata 2009] or variational methods [Blet *et al.* 2003].

Second, LDA lacks an explicit sparcification procedure on the inferred representations. Although we can adjust α to make $\tilde{\boldsymbol{\theta}}$ concentrate much of its mass on a small number of topics *a priori*, it only indirectly influences the sparsity of inferred posterior representations [Ganchev *et al.* 2010]. In practice, using a Dirichlet prior is ineffective in controlling the posterior sparsity of LDA. Fig. 1 shows the sparsity ratio of word code (i.e., number of zeros in the code divided by topic number K) and classification accuracy with different pre-specified Dirichlet parameter α of LDA using variational inference①. We

① In theory, variational methods don't produce zero code elements because of the exponential update rule. But in practice, it is safe to truncate very small values to be zero. Similarly, sampling methods don't have a direct control on the posterior sparisty either.

此为试读,需要完整PDF请访问: www.ertongbook.com