基于深度学习的手机应用分类研究

王超，张华，秦佳伟
（北京邮电大学网络与交换技术国家重点实验室，北京 100876）

摘要：传统的手机应用分类无法分析中文复杂上下文环境，同时随着样本数量增加会发生过拟合问题，针对目前存在问题，本论文研究了深度学习网络，通过构建 DBN 模型使用 CD-K 训练算法学习中文手机应用元数据并进行分类。论文分别使用支持向量机，朴素贝叶斯以及随机森林分类模型做对比，实验结果表明，在手机应用分类方面 DBN 分类结果准确率可达到 82.42%，较其他模型有着更好的表现。

关键词：人工智能；手机应用；深度学习；数据分类
中图分类号：TP181

Classification of mobile applications based on deep learning

WANG Chao, ZHANG Hua, QIN Jiawei
(*Beijing University of Posts and Telecommunications, State Key Laboratory of Networking and Switching Technology, Beijing* 100876)

Abstract: Traditional classification of mobile applications cannot be analyze Chinese complex context, and with the increase in the number of samples will happen over fitting problem, for the existing problems, this thesis studies the deep learning network, constructed by using CD-K training algorithm for learning metadata of applications. This thesis uses support vector machine, naïve Bayes and random forest classification model to compare, the experimental results show that, in mobile application classification, the accuracy rate of DBN classification can reach 82.42%, better than other models.

Keywords: artificial intelligence; mobile applications; deep learning; data classification

0 引言

目前国内手机应用市场超过 300 家，但是不同应用渠道之间并没有严格的分类标准管理手机应用[1]。手机应用的分类研究正是基于该背景下开展的课题，也是对手机应用市场进行细分的重要内容之一。手机应用分类不仅仅对用户有着重要作用，同时对手机厂商以及相关监管部门也有着重要的意义[2]。直观地，手机应用分类研究可以帮助用户了解自己的使用偏好[3];针对手机厂商可以挖掘出手机应用周边的市场潜力，进行有效市场投放;监管部门则可以通过分类数据了解手机应用市场的发展状况[4]。

手机应用中元数据处理方式复杂，尤其对于中文元数据处理，通常进行数据处理以及分类会使用普通的线性模型以及概率模型[5]。对于中文语义关系复杂的语料，普通分类模型无法很好处理词与词之间关系，常常会出现识别率低以及过拟合现象[6]，针对这些问题，本论文根据语料特征构建深度学习结构，处理关系复杂的中文语料[7]。

DBN（deep belief network，深度置信网络）由多层构成，广义角度可以分为显示神经元以及隐藏神经元，显示神经元负责数据样本的输入，而隐藏神经元则进行样本特征的学习[8]。单层神经元由 RBM（restricted boltzmann machines，受限玻尔兹曼机）构成，RBM 是基于玻尔兹曼分布提出网络模型，在 BM（boltzmann machines，玻尔兹曼机）的基础之上取消了层内连接权值，进而转换为二分图，该模型由显示层以及隐藏层构成[9]。通过训练使得模型

作者简介：王超(1992)，男，人工智能与智能信息处理
通信联系人：张华(1978)，女，副教授、博导，密码协议、物联网和云计算安全. E-mail: zhanghua_288@bupt.edu.cn

的整体能量函数降低，此时模型的输出可以很好刻画模型的输入，通过控制模型的输出节点数量可以对原有数据进行特征选取[10]。

论文构建深度学习网络模型对手机应用元数据进行处理，同时使用传统分类方法进行实验包括支持向量机、朴素贝叶斯[11]以及随机森林[12]。通过对比不同分类实验结果证明深度网络结构更适合处理中文手机应用元数据。

1 受限波尔兹曼机以及 DBN 训练

DBN（deep belief network，深度置信网络）是一种非监督的神经网络模型[13]，可以用于数据特征提取、数据分类[14]、图像检索、自动语音识别以及时间序列等任务中[15]。DBN 模型由多层网络节点构成，在训练过程中会对每一层网络单独进行训练使得单层网络达到收敛状态[16]。本论文中构建 DBN 单层网络结构使用的是 RBM，RBM 是在玻尔兹曼机的基础之上进行分层，取消了层内的连接，进而可以用于数据分类以及数据降维，同时可以加速学习速度。

BM 是基于物理现象提出的一种网络模型，模型模拟了物理分子间的运动，可以很好的表示变量之间的高阶作用，同时模型基于一种能量函数，能量函数可以衡量模型的能量大小，通过不断的测量系统能量函数值的大小以及计算误差进行权重调整，使得单层模型最终达到收敛效果[17]，图 1 为 RBM 结构图。

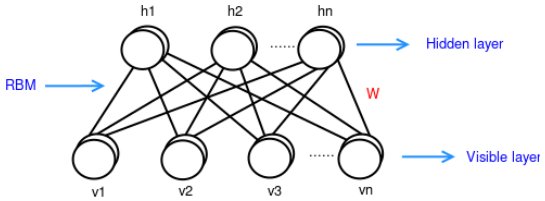

图 1 RBM 模型结构

Fig. 1 RBM model structure

RBM 分为显示单元以及隐藏单元，隐藏单元的目的在于传输中间状态，可以限制隐藏单元数量达到数据降维以及特征抽取的目的，同时增加隐藏层数可以学习变量之间复杂的依赖关系。

计算单层 RBM 系统能量，其中 v 以及 h 代表显示层以及隐藏层单元值，a 以及 b 代表显示层以及隐藏层偏置，w 代表层间连接权值，计算方法见公式 1[8]。

$$E(v,h) = -\sum_{1}^{n} a_i v_i - \sum_{1}^{m} b_j h_j - \sum v_i h_j w_{ij} \quad (1)$$

计算隐藏层以及显示层的单元值，需要根据联合概率密度进行采样，隐藏层以及显示层的联合概率密度计算见公式 2[8]。

$$p(v,h) = \frac{1}{Z} e^{-E(v,h)} \quad (2)$$

计算显示层的概率分布见公式 3[8]。

$$p(v) = \frac{1}{Z} \sum e^{-E(v,h)} \quad (3)$$

模型迭代过程中需要对权值矩阵 w 进行调整，学习规则见公式 4，其中尖括号表示该分配下标分布的期望值[8]。

$$\Delta w_{ij} = \eta(< v_i h_j >_{data} - < v_I h_j >_{model}) \quad 4$$

RBM 采用贪婪训练方式进行训练，过程如下[8]：

RBM 训练过程

初始化阶段：

1. 获得样本输入
2. 选择训练周期 J，学习速率 η 以及参数 k 值
3. 指定显示层、隐藏层数量以及每一层单元数目 n_v，n_h
4. 初始化每一层的偏置向量 a，b 以及连接的权值矩阵 w

训练阶段：

for iter = 1，2，3……，J do

{

调用 $CDK(k, S, RBM(w, a, b); \Delta w, \Delta a, \Delta b)$，生成 $\Delta w, \Delta a, \Delta b$

参数调整：$w = w + \eta(\frac{1}{n_s}\Delta w), a = a + \eta(\frac{1}{n_s}\Delta a), b = b + \eta(\frac{1}{n_s}\Delta b)$

}

通过 RBM 构建 DBN 学习结构，采用逐层训练方式学习手机应用元数据，通过 DBN 提取出中文特征之间的复杂关系。使用提取出的特征进行分类，可以提高手机应用分类的结果，图 2 为构建 DBN 分类模型。

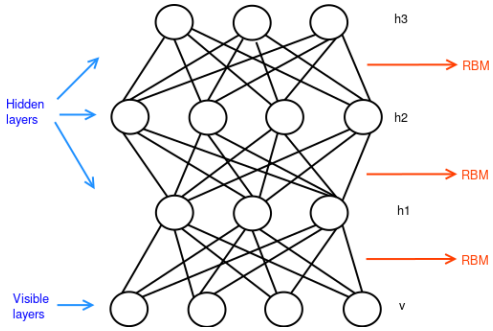

图 2　DBN 分类模型

Fig. 2　DBN classification model

处理后的手机应用元数据信息作为模型输入，每一层构建 RBM 模型，通过能量函数调整 RBM 显示层以及隐藏层之间的权值，层间进行单独训练，每一层的输出是下一层的输入，通过模型训练不断学习中文元数据信息，最终将元数据信息进行分类，获得对应元数据的应用类别。

2　实验分析

2.1　数据样本

手机应用分类属于监督学习方法，收集的样本数据需要进行人工手动标注。元数据收集从目前主流的 300 家国内渠道进行获取，通过网络爬取的方式获得 13000 个元数据信息，不同的渠道中对各个手机应用的分类标准不同，因此需要人工类标注，按照表 1 标准进行手机应用类别标注。

表 1　手机应用分类标准

Tab. 1 Mobile phone application classification standard

类别	类别 ID	内容
网络类	0000	浏览器
商业类	1000	理财，备忘，手机支付，金融理财
通讯类	2000	通信，邮箱，聊天
游戏类	3000	游戏，游戏攻略
多媒体类	4000	影音视频， 网络电视，视频播放器
导航类	5000	旅游出行，地图导航
社交类	6000	社交，博客，微薄
系统类	7000	系统工具，日历，拍摄优化，照相
新闻阅读类	8000	资讯阅读，新闻，漫画，电子书，笑话，字典

收集样本中包含有部分非中文语系手机应用，去除无关元数据，总计 9895 条可用记录，表 2 为手机样本数量分布。

表 2 不同类型元数据样本数量分布

Tab. 2 Metadata sample size distribution

类别 ID	类别	样本数量/个
0000	网络类	493
1000	商业类	812
2000	通讯类	453
3000	游戏类	2429
4000	多媒体类	1986
5000	导航类	487
6000	社交类	416
7000	系统类	1376
8000	新闻阅读类	1443

不同类型的手机应用数量在渠道中分布数量不相同。实验通过网络方式获取元数据信息，根据获取的数据统计发现，游戏类的手机应用较多，有 2000 多个，其次影音类的应用数量也较为丰富，然而社交以及通信邮箱类别的应用较少，这种分布也符合了人们日常使用的行为习惯。

在生活中，很多应用偏向于游戏或者系统工具，对于邮箱类别应用大多采用 PC 端进行访问使用，因此在很多应用渠道中应用类别也呈现出了明显数量上的差异。图 3 为各个类别应用样本分布。

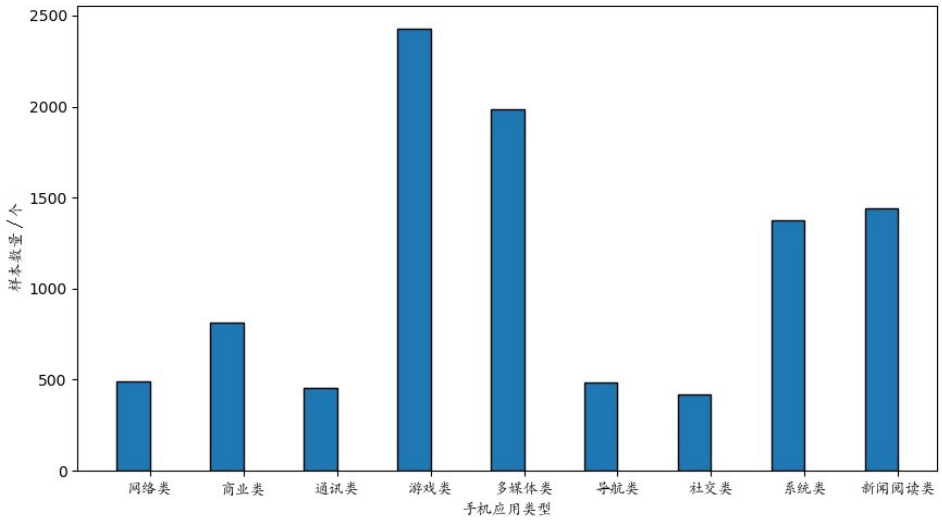

图 3　手机应用样本数量分布

Fig. 3　Sample distribution of mobile applications

实验过程中将 1 万数据分为 10 份样本，样本数量从 1 千到 1 万依次递增 1 千数量，每份样本中随机抽取 10%样本作为测试样本。使用空间向量模型将样本映射成为空间中向量，使用论文构建的深度学习模型以及训练算法进行处理分类。

2.2　分类方法对比

实验使用论文提出的方式进行分类，同时分别采用支持向量机、朴素贝叶斯以及随机森林分类器作为对比，实验结果准确率如图 4 所示。

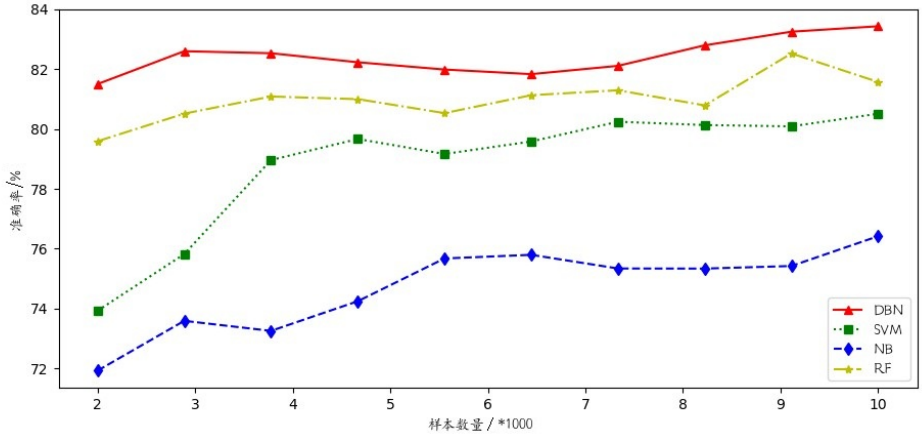

图 4　DBN 分类结果对比图

Fig. 4　Comparison chart of DBN classification results

实验中红色实线为论文采用的 DBN 进行学习分类的实验结果，黄色虚线为随机森林实验结果，绿色虚线为支持向量机实验结果，蓝色虚线为朴素贝叶斯实验结果。分析实验结果，在样本 1k 到 4k 区间各个模型的准确率都有一定的提高，其中支持向量机对于小样本比较敏感，其准确率从 73.93%增长为 79.43%，浮动 5.5%，而 DBN 实验结果从 81.50%提高到 82.45%，浮动仅有 0.95%，在此区间 DBN 分类的准确率平均达到了 82.19%，高于支持向量机、朴素贝叶斯以及随机森林的平均准确率；随着样本数量增加，DBN 分类准确率一直保持在较高

区间，且浮动范围较小。因此可以看出文论采用的 DBN 方法对于手机应用分类准确率有着更好的表现，同时模型的稳定性也更加优秀。

3 结论

针对复杂中文手机应用元数据，论文使用深度学习网络模型，通过分析手机应用元数据特征构建 DBN 学习结构。实验结果表明，深度学习结构分类平均准确率可达 82.42%，较朴素贝叶斯平均准确率提高 7.73%，相比普通线性模型有较大的提高空间，同时分类结果的稳定性也有了一定提升，因此可得出论文构建的深度模型可以更好应用于中文手机应用分类。

[参考文献] (References)

[1] Afzal, SAR Zaidi, MZ Shakir, MA Imran, M Ghogho. The Cognitive Internet of Things: A Unified Perspective[J]. Mobile Netw Appl20, 2015, 20 (1): 72-85.
[2] Al-Fuqaha, A., Guizani, M., Mohammadi, M., Aledhari, M. & Ayyash, M. Internet of Things: A Survey on Enabling Technologies, Protocols, and Applications[J]. Ieee Commun Surv Tut17, 2015, 17 (4): 2347-2376.
[3] Chen, T. S., Chiu, P. S., Huang, Y. M. & Hsieh, W. J. The Effect of Context-Aware Mobile Learning on Chinese Rhetoric Ability of Elementary School Students[J]. J Internet Technol17, 2016, 17 (7): 1309-1316.
[4] Seto, J., Wang, Y. & Lin, X. D. User-Habit-Oriented Authentication Model: Toward Secure, User-Friendly Authentication for Mobile Devices[J]. Ieee T Emerg Top Com3, 2015, 3 (1): 107-118.
[5] Ai, M., Su, X. & Riekki, J. Semantic Reasoning for Context-Aware Internet of Things Applications[J]. Ieee Internet Things4, 2017, 4 (2): 461-473.
[6] Pavlinek, M. & Podgorelec, V. Text classification method based on self-training and LDA topic models[J]. Expert Syst Appl80, 2017, 80: 83-93.
[7] Niazmardi, S., Safari, A. & Homayouni, S. A Novel Multiple Kernel Learning Framework for Multiple Feature Classification[J]. Ieee J-Stars10, 2017, PP (99): 3734-3743.
[8] Cui, Z. Y., Ge, S. S., Cao, Z. J., Yang, J. Y. & Ren, H. L. Analysis of Different Sparsity Methods in Constrained RBM for Sparse Representation in Cognitive Robotic Perception[J]. J Intell Robot Syst80, 2015, 80 (1): S121-S132.
[9] Pang, S. & Yang, X. Y. Deep Convolutional Extreme Learning Machine and Its Application in Handwritten Digit Classification[J]. Comput Intel Neurosc, 2016, 2016 (3): 1-10.
[10] Cheng, C. L., Wang, S., Chen, X. G. & Yang, Y. Y. A Multilayer Improved RBM Network Based Image Compression Method in Wireless Sensor Networks[J]. Int J Distrib Sens N,2016 ,2016 (2):15.
[11] Diab, D. M. & El Hindi, K. M. Using differential evolution for fine tuning naive Bayesian classifiers and its application for text classification[J]. Appl Soft Comput54, 2016, 54: 183-199.
[12] Bellet, A., Bernabeu, J. F., Habrard, A. & Sebban, M. Learning discriminative tree edit similarities for linear classification-Application to melody recognition[J]. Neurocomputing214, 2016, 214 (C):155-161.
[13] Zhu, H. S., Chen, E. H., Xiong, H., Cao, H. H. & Tian, J. L. Mobile App Classification with Enriched Contextual Information[J]. Ieee T Mobile Comput13, 2014, 13(7): 1550-1563.
[14] Saleh, A. I., Al Rahmawy, M. F. & Abulwafa, A. E. A semantic based Web page classification strategy using multi-layered domain ontology[J]. World Wide Web20, 2017,20: 1-55.
[15] Miyoshi, T. & Joichi, H. Comparison with fuzzy reasoning and modified TF-IDF in page grouping for the result of Web retrieval[J]. Int J Innov Comput I3, 2007, 3(2): 307-317.
[16] Azar, A. T., Inbarani, H. H. & Devi, K. R. Improved dominance rough set-based classification system[J]. Neural Comput Appl28, 2017, 28(8): 2231-2246.
[17] Ghahabi, O. & Hernando, J. Restricted Boltzmann machines for vector representation of speech in speaker recognition[J]. Comput Speech Lang47, 2018, 47: 16-29.