

一种适用于卷积神经网络的 Stacking 算法

张笑铭^{1,2}, 王志君², 梁利平²

(1. 中国科学院大学 电子电气与通信工程学院, 北京 100049; 2. 中国科学院微电子研究所, 北京 100029)

摘 要: 为提高卷积神经网络的分类精度, 提出一种结合多个网络的改进 Stacking 算法。将卷积神经网络作为基分类器对数据进行分类, 得到新的样本再经过元分类器分类。为降低元分类器输入数据的维度和多个网络分类结果之间的相关性, 采用主成分分析方法对基分类器的输出进行降维。在数据集上进行分类精度对比实验, 结果表明, 与传统 Stacking、基于平均后验概率算法和基于类投票算法相比, 该算法在同类型网络 and 不同类型网络中, 分类精度均较高且更具有稳定性。

关键词: 卷积神经网络; Stacking 算法; 主成分分析; 降维; 网络结构; 分类精度

中文引用格式: 张笑铭, 王志君, 梁利平. 一种适用于卷积神经网络的 Stacking 算法[J]. 计算机工程, 2018, 44(4): 243-247.

英文引用格式: ZHANG Xiaoming, WANG Zhijun, LIANG Liping. A Stacking Algorithm for Convolution Neural Network[J]. Computer Engineering, 2018, 44(4): 243-247.

A Stacking Algorithm for Convolution Neural Network

ZHANG Xiaoming^{1,2}, WANG Zhijun², LIANG Liping²

(1. School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 100049, China;

2. Institute of Microelectronics of Chinese Academy of Sciences, Beijing 100029, China)

[Abstract] In order to improve the classification accuracy of convolution neural network, an improved Stacking algorithm combining multiple convolution neural networks is proposed. The convolution neural network is used as the base classifier to classify the data, and the new sample is then classified by the meta-classifier. In order to reduce the dimension and correlation of the input data in meta-classifier, the dimension of output data of base classifier are reduced by Principal Component Analysis (PCA). Experimental results on classification accuracy of data sets show that, compared with traditional Stacking, average posteriori probability based algorithm and class voting based algorithm, the classification accuracy of the proposed algorithm is higher and more stable in similar networks and different types of networks.

[Key words] convolution neural network; Stacking algorithm; Principal Component Analysis (PCA); dimensionality reduction; network structure; classification accuracy

DOI: 10.3969/j.issn.1000-3428.2018.04.039

0 概述

随着深度学习技术的发展, 分类问题开始采用深度神经网络结构, 其中卷积神经网络作为一个重要组成成份, 已经应用于计算机视觉等领域。目前, 在卷积神经网络的研究领域, 有多种网络结构被提出, 如 LeNet^[1]、AlexNet^[2]、VGG^[3]、GoogLeNet^[4]、NIN^[5] 和 ResNet^[6] 等, 虽然其效果相比传统分类算法具有明显的提升, 但在对分类准确率有较高要求的场合中, 该单一网络结构只有通过加深层数才能实现精度的提升, 而加深网络的层数会导致训练过程难以收敛, 并产生梯度消失的现象。

为了在现有网络结构下继续提升精度, 学者考

虑将多种结构的卷积神经网络进行结合, 使用 Stacking 算法^[7] 将卷积神经网络当作基分类器, 然后选择合适的元分类器, 最终实现更高的精度。但是, 由于卷积神经网络最终的分种类别较多, 其中大型数据集有 100 多个类别, 此时结合 N 个卷积神经网络的 Stacking 算法, 其输入元分类器的向量维度为 $N \times K$ (K 为类别个数), 该向量维度导致元分类器难以训练以及其泛化能力的降低。文献[8] 提出基于概率的 Stacking 算法, 将所有基分类器的概率输出取平均后作为元分类器的输入。文献[9] 提出基于投票的 Stacking 算法, 选择基分类器概率最大的输出作为分类结果, 将所有结果相加后输入元分类器。上述算法都基于单个样本的特性, 没有考虑整体训

作者简介: 张笑铭 (1992—), 男, 硕士研究生, 主研方向为计算机视觉、图像识别; 王志君, 副研究员、博士; 梁利平, 研究员、博士。

收稿日期: 2017-03-23 **修回日期:** 2017-04-28 **E-mail:** zhangxiaoming@ime.ac.cn

训练样本的统计特性,因而提升效果有限,甚至在某些数据集下出现效果下降的现象。

针对由基分类器输出的维度过高所导致的元分类器泛化能力下降问题,本文提出适用于卷积神经网络的改进 Stacking 算法,在类别数较多的分类任务中利用训练样本的统计特性完成中间数据的降维,以实现精度的提升。

1 Stacking 算法

1.1 基于分类器输出的 Stacking 算法

设由训练样本组成的数据集 D 上有 N 种学习算法,该 N 种学习算法得到 N 个分类器,称为基分类器。数据集 D 中的每个样本经过该 N 个分类器会输出 N 个结果,将该 N 个结果当作一个新的样本,样本标签为数据集 D 中对应的样本标签,使用新的样本训练一个分类器,该分类器称为元分类器。图 1 所示为 Stacking 算法结构框架。使用 Stacking 算法时通过 N 个基分类器将数据进行分类,将得到的 N 个结果组成新的向量再经过元分类器分类后得到最终结果。

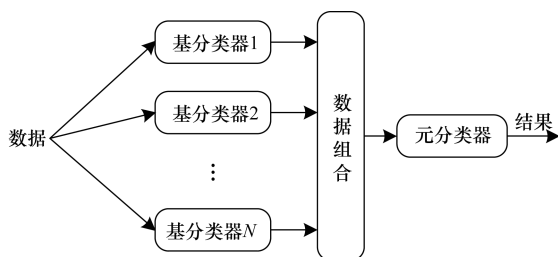


图 1 Stacking 算法结构框架

1.2 基于平均后验概率的 Stacking 算法

基于平均后验概率的 Stacking 算法,其目的是解决由于基分类器个数或任务类别数过多导致的输入元分类器数据维度过高的问题。设有 N 个基分类器, K 个分类类别, x 为输入数据,其中第 i 个基分类器对应的分类结果为 $C_i(x) = [P_{i1}, P_{i2}, \dots, P_{iK}]$, 向量中的元素表示每个类别的概率。对所有分类概率求平均,将其作为元分类器的输入,即

$$[(P_{11} + P_{21} + \dots + P_{N1})/N, (P_{12} + P_{22} + \dots + P_{N2})/N, \dots, (P_{1K} + P_{2K} + \dots + P_{NK})/N]$$

通过该操作,数据从 $N \times K$ 维降为 K 维,但是,该方法掩盖了每个基分类器的预测结果,因此,在类别数较少的数据集中,该方法的精度往往低于基于分类器输出的 Stacking 算法。

1.3 基于类投票的 Stacking 算法

基于类投票的 Stacking 算法将每个基分类器中输出概率最大的类设为 1,其他设为 0,将所有基分

类器的结果相加,用其训练元分类器,此时元分类器的输入同样只有 K 个维度。相比基于平均后验概率的 Stacking 算法,该算法保留了每个基分类器的预测结果,但是忽略了每个基分类器对其他类的预测情况,因此,该算法对基分类器的选择和数据集的变化较敏感。

2 改进 Stacking 算法

2.1 卷积神经网络

卷积神经网络通过不同的卷积核提取图像的局部特征,将该局部特征组合成更高层的全局特征,从而完成图像从局部到整体的映射过程。卷积核特征的提取利用了图像的局部相关性,局部相关性指某像素点的周围点像素值有很大概率是与该点的像素值相同,其中包含了很多冗余信息,而卷积的作用就是过滤掉这些冗余从而得到真正有用的信息。图 2 所示为卷积神经网络结构,其层类型主要有卷积层、池化层、全连接层等。

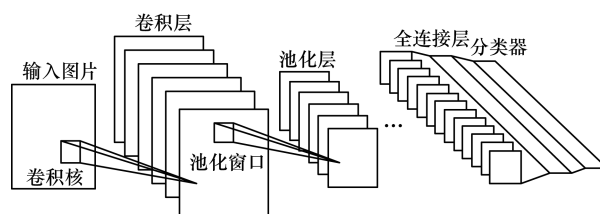


图 2 卷积神经网络结构

卷积神经网络的初始层用来提取底层特征。随着层数的加深,底层特征组合成具有更高视野的特征,即越深的层其感受野范围越大,直至倒数第 2 层,该层获得图片的整体抽象特征。倒数第 1 层是分类器,该层根据前面各层所提取的特征判别图片的类别。

卷积层的作用是提取特征,池化层的作用是扩大局部感受野和减少参数,全连接层的作用是获得整体抽象特征。目前研究中已经提出多种卷积神经网络结构,这些网络结构的不同之处主要在于层的类型以及层与层之间的连接关系。

2.2 主成分分析

主成分分析^[10-12] (Principal Component Analysis, PCA) 是一种常用的降维技术,其将数据的坐标轴进行旋转,选择数据方差最大的方向作为第 1 个新坐标轴,因为方差越大说明信息量越大;选择与第 1 个坐标轴正交的方向作为第 2 个坐标轴,该方向的剩余信息量最大;选择与前 2 个坐标轴都相交的方向作为第 3 个坐标轴。然后,按照上述方法依次选择

其余坐标轴。选择所需数量的坐标轴后,即完成了数据从原始空间到降维后空间的变换。其中,数据方差最大的方向可以通过计算数据的协方差矩阵来获得。对协方差矩阵进行特征值分解,选择最大特征值对应的特征向量作为第1个新坐标轴的方向,选择次大的特征值对应的特征向量作为第2个新坐标轴的方向,……,依此类推,直至完成最后一个坐标轴的选择。

2.3 PCA-Stacking 算法

本文将卷积神经网络作为基分类器,结合主成份分析方法,提出一种改进的 Stacking 算法:PCA-Stacking 算法。根据基分类器具体类型的不同,采用2种组合方式:一种是网络结构相同但构成基分类器的卷积神经网络层数不同,如图3所示;另一种是构成基分类器的卷积神经网络结构完全不同,如图4所示。

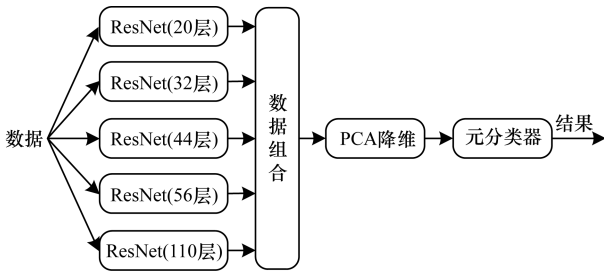


图3 同类型网络结构的 PCA-Stacking 算法结构框架

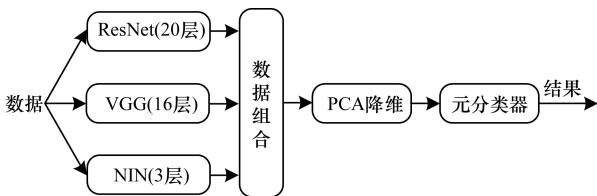


图4 不同类型网络结构的 PCA-Stacking 算法结构框架

多个卷积神经网络对每个类别的分类概率之间相互关联,而 PCA 通过消除数据之间的相关性来达到降维的目的。假设数据集中一共有 K 个类别,对于基分类器的输出也只有 K 个维度是不相关的,因此,PCA 保留的特征向量个数为 K 。

元分类器选择逻辑回归^[13]算法,同时为了满足多分类任务的要求,其训练多个逻辑回归模型,每个模型只对一个类别进行分类,元分类器训练过程中参数的更新选择随机梯度下降法^[14-15]。深度学习的样本数目较大,为了加快训练的收敛速度,并且要逼近全局最优,每次只使用一个样本进行参数的更新。

元分类器的训练过程如下:

步骤1 使用数据集 D 训练多个卷积神经网络

得到在数据集 D 上的 N 个模型,该 N 个模型对数据集 D 进行分类,并将分类结果组成新的样本集 M 。表达式如下: $D = \{(x_i, y_i), i = 1, 2, \dots, m\}$

$$model = \{f_j(x), j = 1, 2, \dots, N\}$$

$$M = \{((f_1(x_i), f_2(x_i), \dots, f_N(x_i)), y_i), i = 1, 2, \dots, m\}$$

步骤2 将样本集 M 组合成矩阵 X 的形式, X 的每一行为一个样本向量,列数等于样本个数。按式(1)计算样本集 M 的协方差矩阵 C 。

$$C = \frac{1}{m-1} X^T X \quad (1)$$

步骤3 计算协方差矩阵 C 的特征值和特征向量,将特征值从大到小进行排序,保留前 K (K 为数据集类别个数)个特征值和其对应的特征向量。该 K 个特征向量组合成转换矩阵 T 。

步骤4 对样本集 M 中的每个样本使用转换矩阵 T 进行降维,得到降维后的样本集 S ,表达式如下:

$$S = \{((S_{i1}, S_{i2}, \dots, S_{iK}), y_i), i = 1, 2, \dots, m\} \quad (2)$$

步骤5 用样本集 S 训练元分类器,其中通过交叉验证法选择合适的元分类器参数。

数据经过基分类器的分类后,测试过程将该分类结果组合成新的样本集,训练过程的转换矩阵 T 对该样本集进行降维,降维数据经过元分类器的分类后,得到最终测试结果。

3 实验结果与分析

3.1 PCA 维度个数保留

PCA 对数据进行降维时会伴有信息量的损失,用保留特征值占有所有特征值的比例(方差保留百分比)衡量该损失的大小,表达式如下:

$$r = \frac{\sum_{j=1}^K \lambda_j}{\sum_{j=1}^{N \times K} \lambda_j} \quad (3)$$

r 越大表示保留的信息量越大,即损失越小。

本文在 CIFAR-10 和 CIFAR-100 2 个数据集上进行验证实验。CIFAR-10 数据集共有 10 个类别,分为训练集和测试集,训练集中每个类别有 5 000 张图片,测试集中每个类别有 1 000 张图片。CIFAR-100 数据集共有 100 个类别,训练集中每个类别有 500 张图片,测试集中每个类别有 100 张图片。记录方差保留百分比 r 随特征值保留个数的变化情况,实验结果如图 5 所示。从图 5 中可以看出,在同类型网络结构中,当特征值保留个数与数据集类别数相同时,在 CIFAR-10 数据集上保留了 99.9% 的方

差,在 CIFAR-100 数据集上保留了 97.6% 的方差;在不同类型的网络结构中,当特征值保留个数与数据集类别数相同时,在 CIFAR-10 数据集上保留了 99.8% 的方差,在 CIFAR-100 数据集上保留了

96.8% 的方差。通过图 5 还可以看出,继续增加特征值保留个数时,方差保留百分比的增加有限,因此,当特征值的保留个数等于类别数时,可以在实现数据压缩的同时保留足够的信息量。

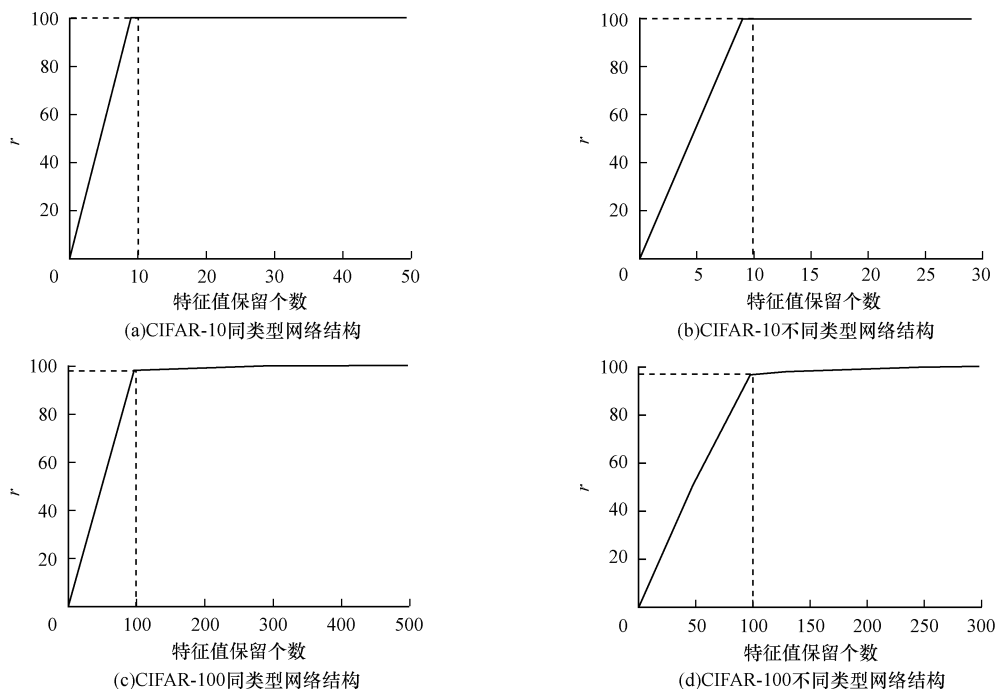


图 5 方差保留百分比与特征值保留个数的关系

3.2 网络结构对比

使用 Caffe 深度学习框架测试结构为 ResNet、VGG 和 NIN 的卷积神经网络,在 CIFAR-10 和 CIFAR-100 数据集上进行分类精度实验。然后,将所得到的模型作为基分类器,实现基于分类器输出 (Stacking)、基于平均后验概率 (Astacking) 和基于类投票 (Vstacking) 的对比算法,并与本文算法 (PCA-Stacking) 进行比较。

3.2.1 同类型网络

本文实验中同类型网络选择 ResNet 结构,训练层数分别为 20、32、44、56 及 110 的模型,精度测试结果如表 1 所示。

表 1 不同层数结构分类精度结果 %

数据集	20 层	32 层	44 层	56 层	110 层
CIFAR-10	91.24	91.97	92.73	92.81	93.17
CIFAR-100	65.57	66.63	66.67	68.60	70.37

从表 1 中可以看出,随着网络层数的加深,分类精度逐渐提高,但精度提升的难度越来越大。在 CIFAR-10 数据集中,110 层的 ResNet 比 56 层的 ResNet 分类精度提升了 0.36%。在 CIFAR-100 数据集中,110 层的 ResNet 比 56 层的 ResNet 分类精度提升了 1.77%。

结合多个网络时不同算法的分类精度对比结果如表 2 所示。

表 2 同类型网络不同算法分类精度对比 %

数据集	Stacking 算法	Astacking 算法	Vstacking 算法	PCA-Stacking 算法
CIFAR-10	94.41	94.61	94.37	94.63
CIFAR-100	74.71	75.12	70.56	75.29

从表 2 可以看出,在 CIFAR-10 和 CIFAR-100 数据集上,所有算法的分类精度都高于单一网络结构,其中,本文算法在 CIFAR-10 数据集上比单一网络结构最高精度提升了 1.46%,在 CIFAR-100 数据集上比单一网络结构最高精度提升了 4.92%,同时,该算法的分类精度也高于其他对比算法。

3.2.2 不同类型网络

不同类型网络选择 20 层的 ResNet、16 层的 VGG 和 3 层的 NIN 进行实验,精度测试结果如表 3 所示。

表 3 不同网络结构分类精度对比结果 %

数据集	ResNet-20 结构	VGG-16 结构	NIN-3 结构
CIFAR-10	91.24	90.55	89.90
CIFAR-100	65.57	62.42	64.01

从表3可以看出,不同网络结构的分类精度比较接近,该设计消除了单个网络精度过低或过高对算法效果造成的影响,结合多个不同结构的分类精度对比结果如表4所示。

表4 结合不同结构的分类精度对比 %

数据集	Stacking 算法	Astacking 算法	Vstacking 算法	PCA-Stacking 算法
CIFAR-10	91.20	92.41	92.34	92.78
CIFAR-100	63.61	69.86	68.56	70.29

从表4可以看出,在不同网络结构下,Stacking 算法分类精度低于单一网络结构,原因在于不同网络结构下,各个网络之间的相关性较弱,造成元分类器泛化能力下降,因此,其分类精度低于单一网络结构。本文 PCA-Stacking 算法使用了样本的统计特性,因此,其分类精度比较稳定,在 CIFAR-10 数据集上比单一网络结构最高精度提升了 1.54%,在 CIFAR-100 数据集上比单一网络结构最高精度提升了 4.72%。

4 结束语

本文提出一种适用于卷积神经网络的改进 Stacking 算法,该算法使用 PCA 对多个卷积神经网络的分类结果进行降维,将降维后的数据输入元分类器。实验结果表明,与其他 Stacking 算法相比,该算法利用了样本的统计特性,提升了分类精度,且在不同数据集下精度的提升较稳定。下一步将研究分类精度与卷积神经网络组合方式的关系,以及元分类器的类型对该算法的影响。

参考文献

- [1] LECUN Y, BOSE B E, DENKER J S. Backpropagation applied to handwritten zip code recognition[J]. Neural Computation, 1989, 1(4): 541-551.
- [2] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks[C]// Proceedings of the 25th International Conference on Neural Information Processing Systems. [S.l.]: Curran Associates, Inc., 2012: 1097-1105.
- [3] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[EB/OL]. [2017-03-05]. <http://www.robots.ox.ac.uk/~vgg/publications/2015/Simonyan15/simonyan15.pdf>.
- [4] SZEGEDY C, LIU W, JIA Y. Going deeper with convolutions[C]// Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2015: 1-9.
- [5] LIN M, CHEN Q, YAN S. Network in network[EB/OL]. [2017-03-15]. <https://arxiv.org/pdf/1312.4400.pdf>.
- [6] HE K, ZHANG X, REN S. Deep residual learning for image recognition[C]// Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2016: 770-778.
- [7] WOLPERT D. Stacked generalization[J]. Neural Networks, 1992, 5(2): 241-260.
- [8] TSOUMAKAS G, VLAHAVAS I. Distributed data mining of large classifier ensembles[C]// Proceedings of Companion Volume of the 2nd Hellenic Conference on Artificial Intelligence. Berlin, Germany: Springer, 2002: 249-256.
- [9] 韦艳艳, 李陶深. 一种基于投票的 Stacking 方法[J]. 计算机工程, 2006, 32(7): 199-201.
- [10] MACKIEWICZ A, RATAJCZAK W. Principal components analysis (PCA) [J]. Computers and Geosciences, 1993, 19(3): 303-342.
- [11] 刘丽敏, 樊晓平, 廖志芳. 基于迹范数的 L1-PCA 算法[J]. 计算机工程, 2013, 39(4): 199-202.
- [12] ZOU H, HASTIE T, TIBSHIRANI R. Sparse principal component analysis[J]. Journal of Computational and Graphical Statistics, 2006, 15(2): 265-286.
- [13] KING G, ZENG L. Logistic regression in rare events data[J]. Political Analysis, 2001, 9(2): 137-163.
- [14] ZHANG T. Solving large scale linear prediction problems using stochastic gradient descent algorithms [C]// Proceedings of the 21st International Conference on Machine Learning. New York, USA: ACM Press, 2004: 919-926.
- [15] RECHT B, RE C, WRIGHT S, et al. Hogwild: a lock-free approach to parallelizing stochastic gradient descent[C]// Proceedings of the 24th International Conference on Neural Information Processing Systems. New York, USA: ACM Press, 2011: 693-701.
- [10] BRESEE J S, HECKERMAN D, KADIE C. Empirical analysis of predictive algorithms for collaborative filtering[J]. Uncertainty in Artificial Intelligence, 1998, 98(7): 43-52.
- [11] JASCHKE R, MARINHO L, HOTH O A, et al. Tag recommendations in social bookmarking systems[J]. AI Communications, 2008, 21(4): 231-247.
- [12] KDDCup[EB/OL]. [2012-12-21]. <http://acm.sjtu.edu.cn/courses/kddcup/2012>.
- [13] NANOPOULOS A. Item recommendation in collaborative tagging systems[J]. IEEE Transactions on Systems Man & Cybernetics, Part A: Systems & Humans, 2011, 41(4): 760-771.
- [14] AMANDI A, GODOY D. Hybrid content and tag-based profiles for recommendation in collaborative tagging systems[C]// Proceedings of Latin American Web Conference. Washington D.C., USA: IEEE Press, 2008: 58-65.
- [15] SCHLEE C. Targeted advertising technologies in the ICT space[M]. Berlin, Germany: Springer, 2013.
- [16] CANTADOR I, BELLOGIN A, IGNACIO F T, et al. Semantic contextualisation of social tag-based profiles and item recommendations[J]. E-Commerce and Web Technologies, 2011, 85: 101-113.

编辑 吴云芳

编辑 索书志

(上接第242页)