Hans Journal of Data Mining 数据挖掘, 2020, 10(1), 90-95
Published Online January 2020 in Hans. http://www.hanspub.org/journal/hjdm
https://doi.org/10.12677/hjdm.2020.101009

Hans汉斯

Research on Recognition of Printed Mathematical Formula Based on SVM

Weihai Wen, Lihong Yang, Yao Zhou

School of Mathematics, South China University of Technology, Guangzhou Guangdong
Email: 1107127710@qq.com

Received: Jan. 2nd, 2020; accepted: Jan. 9th, 2020; published: Jan. 16th, 2020

Abstract

Traditional mathematical formula recognition, usually based on OCR technology for image and text recognition, cuts the symbol of the target formula, builds the mathematical symbol database, compares the similarity, and then returns the symbol name of the maximum similarity as the recognition result. In view of the actual situation, there are some differences in the formula, such as font size, thickness, italics, various fonts and so on. Based on the characteristics of printed mathematical formulas, this paper reconstructs the character standard library, and combines with the machine learning idea, uses SVM algorithm to recognize formulas, and further extracts the character features, improves the accuracy of formula recognition. The experimental results show that the recognition results are good.

Keywords

Formula Recognition, Standard Library, Machine Learning, SVM

基于SVM的印刷体数学公式识别的研究

文伟海，杨立洪，周　瑶

华南理工大学数学学院，广东 广州
Email: 1107127710@qq.com

收稿日期：2020年1月2日；录用日期：2020年1月9日；发布日期：2020年1月16日

摘　要

传统的数学公式识别，通常建立在OCR技术进行图片文字识别的基础上，对目标公式进行符号切割，通过构建数学符号数据库，然后两两比较相似度，然后返回最大相似度的符号名称，作为识别结果。该方

文章引用: 文伟海, 杨立洪, 周瑶. 基于 SVM 的印刷体数学公式识别的研究[J]. 数据挖掘, 2020, 10(1): 90-95.
DOI: 10.12677/hjdm.2020.101009

法，对数学符号数据库要求极高，鉴于实际情况，公式存在字号大小、粗细体、正斜体、各种字体等差异，导致该方法识别效果不佳。本文基于印刷体数学公式特点，重新构建字符标准库，并结合机器学习思想，应用SVM算法进行公式识别，并进一步提取字符特征，提升公式识别精度，实验结果显示，识别结果良好。

关键词

公式识别，标准库，机器学习，SVM

1. 引言

随着互联网和信息技术的不断发展，中国在线教育已逐步进入智能教育时代，如拍照搜题，拍照阅卷以及拍照题库等教育类型应用层出不穷。另外，与传统的纸质书籍相比，电子书籍具有便于修改、储存和检索的优势，越来越多的人们倾向于从电子书籍中学习新知识。因此，将印刷体扫描图像转化为可编辑的文本，对于在线教育的发展与科技发展水平、教育理念变革以及用户教育需求升级和生活方式转变具有非常重大的意义。目前这方面发展比较成熟的技术是光学字符识别技术(OCR)，能够较精确地识别中英文以及阿拉伯数字，但对数学公式的识别效果不佳。数学公式符号种类繁多、公式结构复杂，以及符号含义的多样性，让传统的 OCR 技术力所不及。本文将研究提高数学公式识别精度，为数学公式的全面识别提出一点新思路。

数学公式识别作为实用性较强的技术引起了国内外专家和学者的广泛关注和研究。1968 年，Anderson 首次提出数学公式的识别问题[1]。1996 年，Blostein 和 Grbavec 给出了公式识别的定义以及提出了公式识别的重新构图法。在 Okamoto 等人的系统中[2] [3]，首先采用目标结构分析法递归分割字母以及符号，然后建立相对应的字符关系树，最后传统的模板匹配算法来进行数学公式的识别。Lee H J 和 Lee M C 创建的系统中，通过提取数学公式行高度、文档位置信息、相邻行间隔大小等特征[4] [5]，来识别和提取公式。为了解决系统误判问题，采用连通域分割的方法，先切分公式、优化公式粘连、字符识别和逻辑分析重组、最后把结果储存为 LATEX 格式。国内有靳简明的 MathReader 数学公式识别系统，其利用 Parzen 窗进行公式定位，结合水平垂直投影技术、连通域分割技术和统计学特征分析技术进行公式识别，然后定义了 11 种公式来重构表达式并输出[6]。王琪辉则建立了面向公式符号识别的卷积神经网络结构，并通过大量的对比实验确定网络的最优参数[7]。

综上所述，数学公式的识别问题研究较早，但是数学公式(特别是微积分公式)结构复杂，识别难度大，还是有很多亟待解决的难题。本文在结合前人研究的研究成果，通过对数学公式进行分析、总结，进一步提取公式符号特征，使用支持向量机(SVM)对数学公式进行识别，并加入朴素贝叶斯(Naive Bayes, NB)模型作为对比分析。NB 是基于条件概率的分类算法，通过概率大小来进行分类，而 SVM 通过数据点到分割线的距离远近来进行分类。在传统机器学习领域，NB 和 SVM 是最常用分类算法，在不同的分类问题上性能也有所不同。本文选取 NB 作为对比，旨在测试 SVM 模型的效率和准确性，力求建立一个性能优良的 SVM 模型，为传统机器学习模型在公式识别技术的研究提供一些指导。

2. 关键技术

根据结合机器学习的思想，采用 word、latex 常见的数学公式字符作为训练样本，基于支持向量机(SVM)构造数学公式识别分类器，提高公式识别精度。其基本流程图如图 1：

Figure 1. Flow chart of formula recognition
图 1. 公式识别流程图

1) 图像倾斜校正

基于 Hough 变换图像倾角检测方法。对图像边缘线进行 Hough 变换，根据 Hough 变换对图像交点进行投票，找出边缘曲线的倾斜角，并以此矫正图像的倾斜角度。

2) 公式字符切割

基于数学公式特征符号的结构，将公式中的字符分割成独立个体，并保留字符的位置信息。本文结合投影法切割速度快和连通法切割效果好的特点，将两种算法整合，优化公式切割流程，能在保留准确位置信息前提下，提高切割准确率，实验证明该方法的切割效果极佳。

3) 公式字符识别

构建数学公式常用字符的字符库，本文提取了每个字符的“九宫格”和宽高比特征，采用 SVM 建立分类模型。同时，将 SVM 与模板匹配和朴素贝叶斯方法的识别结果进行比较，在速度和精度上，SVM 都胜于上述两种方法。

4) 公式结构分析与组合

本文在公式切割中，通过二值化图像矩阵提取公式字符的位置特征 (x, y, w, h) (其中 x，y 表示该字符在二值化矩阵中的坐标，w，h 分别表示宽和高)，根据数学公式的符号组合方式，构建不同组合逻辑。例如，根号的“半包围”组合方式、定积分的“上下标”组合方式等，对公式进行结构分析。

3. 核心算法

3.1. 公式切割算法

通常公式切割的算法是投影法和连通法，两种算法各有优缺点。投影法算法复杂度低，切割速度快，但是切割效果不佳，如 $\sqrt{b^2-4ac}$ 型的公式无法进一步切割；连通法的特点是，切割效果好，但是算法复杂度相对较高，切割速度慢。

本文将两种算法整合，进行公示切割，实验证明该方法的切割效果极佳。算法步骤如下：

Step 1：对待识别公式行，用投影法进行公式切割，返回切割结果，并记录切割完毕的所有字符的位置特征(x, y, w, h)；
Step 2：识别切割好的字符，删除能识别的字符，并返回识别结果$\{'x':x,'y':y,'w':w,'h':h,'value':value\}$；
Step 3：未能识别的字符，用连通法进行切割；
Step 4：再次识别切割好的字符，返回识别结果$\{'x':x,'y':y,'w':w,'h':h,'value':value\}$。

3.2. 公式识别算法

3.2.1. 模型原理

本文运用支持向量机(SVM)算法进行公式识别，其基本原理[8]如下：

假设训练数据集$S=\{(x_1,y_1),(x_2,y_2),\cdots,(x_m,y_m)\}$，其中$x_i\in R^n$，$y_i\in\{-1,1\}$，$-1$表示负类，1表示正类，一般的二分类问题，为求得最优分类超平面需要求解如下优化问题：

$$\min_{\alpha}\frac{1}{2}\sum_{i=1,j=1}^{m}\alpha_i\alpha_j y_i y_j K(x_i,x_j)-\sum_{i=1}^{m}\alpha_i$$
$$s.t.\ \ 0\le\alpha_i\le C, i=1,2,\cdots,m$$
$$\sum_{i=1}^{m}\alpha_i y_j=0$$

其中，$K(x_i,x_j)$为核函数，α_i为拉格朗日系数，C为惩罚系数。

用SVM进行公式字符识别，其分类为多分类问题，一般为线性不可分，文本中SVM针对线性不可分的分类问题加入核函数。使用径向基内核(Radial basis function kernel, RBF)，其分界面为曲线，对线性不可分问题有良好的拟合效果。

另外，SVM一般应用于二分类问题，本文使用python的机器学习库sklearn中的支持向量机模型svm.svc()，用于本文的字符识别多分类问题。其基本原理是构造多层二分类器，将一个n分类问题分解为第1类和剩余$n-1$类的二分类问题，以此类推，构造最终的多分类模型。

3.2.2. 模型评估

对于分类问题，其模型结果，可用“混淆矩阵”呈现如下表1：

Table 1. Confusion matrix
表 1. 混淆矩阵

真实情况	预测结果	
	正例	反例
正例	*TP*	*FN*
反例	*FP*	*TN*

查全率：

$$p=\frac{TP}{TP+FP}$$

查准率：

$$R=\frac{TP}{TP+FN}$$

F_1-score：

$$F_1 = \frac{2 \times P \times R}{P + R} = \frac{2TP}{\text{样例总数} + TP - TN}$$

4. 实证分析

4.1. 模型数据集

根据印刷体数学公式常见字符，本文构建的数学公式字符库包括 word、latex、pdf 等文档形式中的数字、字母、希腊字母、数学符号等，鉴于图像大小、清晰度不同等因素，选取多个字号构建字符库，具体的字符库构成见下表 2：

Table 2. Composition of character library
表 2. 字符库构成

类别	数字、字母、希腊字母、公式符号
字号	六号、小五号、五号、11 号、小四号、四号、小三号、三号、小二号
字体	宋体、Calibri (西文正文)、微软雅黑、楷体、Times New Roman 等 9 种字体。
样本源	word、pdf、LaTeX、mathtype

4.2. 模型结果

将数据集 3/7 比例随机划分为训练集和测试集，建立模型，返回模型结果如下表 3：

Table 3. Bayes and SVM model accuracy
表 3. Bayes 和 SVM 模型准确率

方法	精度	召回率	F_1-score
Bayes	0.883	0.866	0.868
SVM	0.973	0.977	0.974

测试结果显示，SVM 方法进行字符分类，精度高，识别效果较好。

4.3. 测试结果

为进一步测试模型识别效果，随机抽取三篇英文版数学文档，按照文档识别流程，分别采用直接匹配法、Bayes 法识别、SVM 法识别，统计识别结果如下表 4：

Table 4. Test results of English mathematics documents
表 4. 英文数学文档测试结果

方法	精度	时间
模板匹配法	84%	59 min
Bayes	74%	12 min
SVM	96%	9 min

测试结果显示，SVM 方法进行字符识别，精度高，识别速度快。

5. 结论

本文构建印刷体数学公式字符数据标准库，结合机器学习的思想，构建大量的数学公式字符作为训

练样本，基于 SVM 构造数学公式识别多分类器。采用直接模板匹配、朴素贝叶斯和支持向量机三种方法进行比较，实验结果表明，在传统机器学习领域，支持向量机模型在印刷体数学公式识别中有非常良好的效果。

参考文献

[1] Anderson, R.H. (1968) Syntex-Directed Recognition of Hand-Printed Two-Dimensional Mathematics. In: *Interactive Systems for Experimental Applied Mathematics*. Academic Press, New York, 436-459. https://doi.org/10.1016/B978-0-12-395608-8.50048-7

[2] Twaakyondo, H.M. and Okmoto, M. (1995) Structure Analysis and Recognition of Mathematical Expressions. *Proceedings of the* 3*th International Conference on Document Analysis and Recognition*, Montreal, Canada, 14-16 August 1995, 430-437.

[3] Okamoto, M., Imai, H. and Takagi, K. (2001) Performance Evaluation of a Robust Method for Mathematical Expression Recognition. *Proceedings of the* 6*th International Conference on Document Analysis and Recognition*, Seattle, WA, USA, 13 September 2001, 121-128.

[4] Lee, H.-J. and Lee, M.-C. (1994) Understanding Mathematical Expressions Using Procedure Oriented Transformation. *Pattern Recognition*, **27**, 447-457. https://doi.org/10.1016/0031-3203(94)90121-X

[5] Lee, H.J. and Wang, J.S. (1995) Design of a Mathematical Expression Recognition System. *Proceedings of* 3*rd International Conference on Document Analysis and Recognition*, Montreal, Canada, 14-16 August 1995, 1084-1087.

[6] Scientific, W. (1997) Handbook of Character Recognition and Document Image Analysis. World Scientific, Singapore.

[7] 王琪辉. 基于深度学习的印刷体数学公式符号识别方法研究[D]: [硕士学位论文]. 沈阳: 沈阳工业大学, 2016.

[8] 张学工. 关于统计学习理论与支持向量机[J]. 自动化学报. 2000, 26(1): 32-42.