Hadoop 下自适应随机权值多特征融合图像分类

王 敏[1],陈立潮[1],曹建芳[2],潘理虎[1]
(1. 太原科技大学 计算机科学与技术学院,山西 太原 030024;
2. 忻州师范学院 计算机科学与技术系,山西 忻州 034000)

摘 要:图像具有丰富的语义信息,面对越来越复杂的图像场景,单一特征往往不能准确描述图像内容,因此多特征融合的方式在描述图像中得到了广泛应用。在融合过程中,针对准确确定权值的问题,提出一种自适应的基于随机权值的多特征融合图像分类算法。首先生成随机权值矩阵,然后利用自定义的融合公式,得到融合特征矩阵。为验证算法的效果,将融合后的特征输入 SVM,通过 MapReduce 框架的 Map 过程和 Reduce 过程得到最优权值组合。在 Corel1000 数据集上的实验结果表明,与单特征、1∶1∶1 融合等相比,该算法分类正确率高、运行耗时少,当训练 SVM 的个数达到 120 时,系统加速比几乎呈线性增长的趋势,验证了 Hadoop 平台应对高复杂性算法时的有效性。

关键词:自适应随机权值;特征融合;支持向量机;图像分类;Hadoop 平台
中图分类号:TP391 **文献标识码**:A **文章编号**:1673-629X(2018)11-0030-05
doi:10.3969/j.issn.1673-629X.2018.11.007

Multi-feature Fusion Image Classification of Adaptive Random Weight Based on Hadoop

WANG Min[1],CHEN Li-chao[1],CAO Jian-fang[2],PAN Li-hu[1]
(1. School of Computer Science and Technology,Taiyuan University of Science and Technology,
Taiyuan 030024,China;
2. Department of Computer Science and Technology,Xinzhou Teachers' University,Xinzhou 034000,China)

Abstract:Images possess rich semantic information. In the face of increasingly complex image scenes,single features often cannot accurately describe the image content. Therefore,the method of multi-feature fusion has been widely used in the description of images. In the process of fusion,we propose an adaptive multi-feature fusion image classification algorithm based on random weights to accurately determine the weight. Firstly,it generates a random weight matrix and then obtains a fusion feature matrix by using a self-defined fusion formula. To verify the effect of the algorithm,the fused features are input into SVM to gain the optimal weight combination through the Map process of MapReduce framework and Reduce process. Experiment on the Corel1000 dataset shows that the proposed algorithm has the advantages of high classification accuracy and low running time compared with the single feature,1∶1∶1 fusion and so on. When the number of SVM training reaches 120,the speed-up ratio of the system almost will present linear tendency,which verifies the effectiveness of the Hadoop platform in dealing with high complexity algorithms.

Key words:adaptive random weights;feature fusion;support vector machine;image classification;Hadoop platform

0 引 言

随着互联网技术的变革,大数据技术、深度学习、人工智能等相关研究变得越来越重要,其中大数据技术在各领域应用广泛,它不仅仅用于分析处理海量数据,还用于解决由于计算量大导致算法效率较低的问题[1]。图像分类可以使图像数据得到归并,对于图像的识别和检索都有非常重要的作用。当前,图像分类的主要方法是以图像底层特征(如颜色、纹理、形状等)为基础,利用相关算法训练出分类模型后预测图像的类别信息。

Azhar 等[2]提取图像的 Sift 特征后,用 SVM 构建分类模型,实现对蜡染图像的分类。Camlica 等[3]利用

收稿日期:2017-12-11 **修回日期**:2018-04-25 **网络出版时间**:2018-06-29
基金项目:山西省自然基金项目(2014011019-3);山西省科技重大专项(20121101001);山西省—中科院科技合作项目(20141101001)
作者简介:王 敏(1991-),男,硕士研究生,CCF 会员,研究方向为智能信息处理、大数据技术;陈立潮,博士,教授,研究方向为智能信息处理;曹建芳,博士,教授,研究方向为数字图像理解、大数据技术;潘理虎,博士,副教授,研究方向为智能信息处理。

LBP 特征,通过支持向量机对医学图像进行了分类。由于图像底层特征的提取和表示对图像的分类性能有着重要的影响,一般来说,采用单一特征不能更好地描述一幅图像的内容,因此,从图像中提取多种特征通过一定方法组合成能更好地描述图像内容的特征向量成为很多研究者的选择[4]。但不同特征在图像中的重要性并不完全相同,当采用多个底层特征来描述图像内容时,如果只是将单个特征简单拼接,就不能控制每个特征对分类结果的影响系数,这在一定程度上影响了分类器的准确率。因此,文中提出了一种自适应的随机权值多特征融合图像分类算法(Multi-feature fusion classification algorithm with random weight,MFFRW),并通过实验对其进行验证。

1　随机权值多特征融合分类算法

数据融合可以分为像素级、特征级和决策级三个层次[5]。特征级融合属于中间层,它将来自传感器的原始数据提取特征后进行融合。已有研究成果表明[6-7],多特征融合能够提高图像分类或识别的性能。

特征融合的关键是寻找最优权值组合,因此,设计权值确定算法就变得尤为重要。最初的研究主要是通过反复实验确定,但这种方法受主观因素影响较大,所以很多研究者就提出了自适应的权值确定算法。张春森等[8]依据每个特征对应的分类正确率,利用提出的权值计算公式自适应确定在融合特征中每个特征对应的权值,但权值计算公式中的常量取值由人工确定,受人为经验影响。李玉峰等[9]用预先训练好的单个特征对应的分类器模型分别识别训练集中的数据,如果能正确识别,那么该特征对应的权重加 1,如此循环计算出每个特征对应的权值。该算法虽然避免了人为干预,但需要多次输入输出,算法复杂度高,且预先的分类器模型质量与初始训练样本质量密切相关。袁广林等[10]提出一种基于概率分布可分性判据确定特征融合权重的方法,根据目标与背景特征值的概率分布动态计算它们之间的区分度,克服了利用单一特征跟踪易受相似目标与背景的影响,提高了算法鲁棒性,但该方法只适应于目标跟踪等相关应用场景,对多分类场景并不适用。

综上所述,权值确定受人为影响小、算法复杂度低和融合效果好成为未来研究的方向。张春森、李玉峰等提出的自适应权值算法虽然在一定程度上提高了融合效果,但受算法设计思想、复杂度和应用场景的限定,权值的确定还受主观因素影响,算法复杂度较高,且应用场景单一。因此,提出一种受主观因素影响小、复杂度低、能适应多种场景的自适应权值确定算法就变得非常重要。

1.1　MFFRW 算法框架设计

MFFRW 算法属于特征级融合的图像分类算法,包括特征提取、特征归一化、随机权值矩阵生成、特征融合、训练分类器、得到最优权值组合及对应的分类器 6 个步骤,框架如图 1 所示。

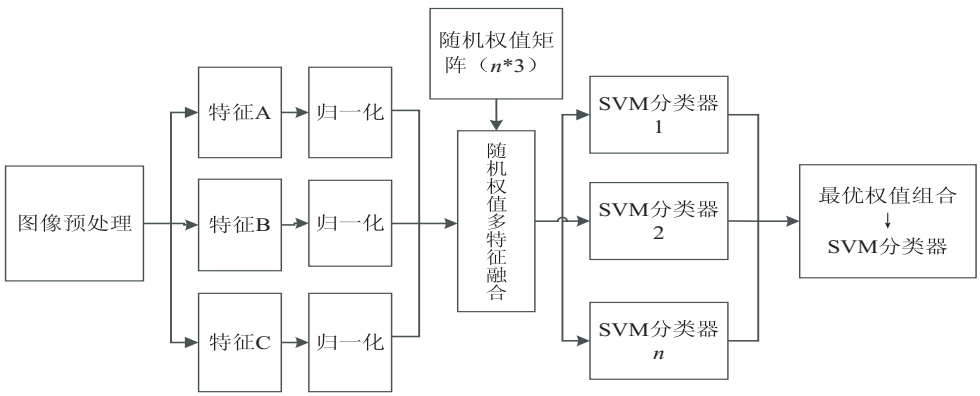

图 1　MFFRW 算法框架

该算法将单一特征通过随机权值矩阵形成融合特征,然后通过训练多个分类器进一步得到最优权值组合。算法在特征提取阶段和 SVM 训练阶段彼此独立,满足并行算法的执行条件,因此可利用大数据技术提高运算效率。

1.2　特征归一化

不同特征向量的量纲不同,取值范围也存在较大差异,当融合成一个特征向量时,需要按照一定的规则加以处理。同一幅图像的不同特征之间,数值上存在很大的悬殊,为了避免其对分类结果的影响,文中采用线性归一化方法将特征值缩放到一个指定范围,用如下方法进行特征归一化。

将原特征向量的值用数组 $\{A_1,A_2,\cdots,A_n$ 表示,p 表示标准化后的值,$A_k \notin \{\max(A_i),\min(A_i)$ 。当 $A_k=\max(A_i)$ 时,$p=1$;$A_k=\min(A_i)$ 时,$p=0$;否则,$p=\dfrac{A_k}{\max(A_i)-\min(A_i)}$。归一化处理可以在很大程度上削弱取值范围差异对不同特征向量重要性的

影响。

1.3 融合规则

为了实现 MFFRW 算法,进一步给出它的数学描述形式:

(1)随机权值矩阵。

设 $\boldsymbol{M}_{n\times 3}=\begin{bmatrix}A_1 & B_1 & C_1\\ A_2 & B_2 & C_2\\ \vdots & \vdots & \vdots\\ A_n & B_n & C_n\end{bmatrix}_{n\times 3}$ 为随机权值矩阵,矩阵 $\boldsymbol{M}_{n\times 3}$ 的秩 $R(M_{n\times 3})=n$。设 m 为 $\boldsymbol{M}_{n\times 3}$ 中任意随机生成的元素,$0<m<1$ 且 m 为 2 位有限小数,并且 $A_i+B_i+C_i=1(i=1,2,\cdots,n)$ 时,生成一组随机权值 $[A_i,B_i,C_i](i=1,2,\cdots,n)$,$n$ 组随机权值组成随机权值矩阵。

(2)融合特征向量矩阵。

用 x,y,z 分别代表 3 种不同的图像底层特征,通过式 1 得到融合特征向量矩阵 $\boldsymbol{W}_{n\times 1}$,矩阵的每一行代表一个融合特征向量,$n$ 的大小取决于随机权值矩阵的行数。

$$\boldsymbol{W}_{n\times 1}=\begin{bmatrix}A_1x+B_1y+C_1z\\ A_2x+B_2y+C_2z\\ \vdots\\ A_nx+B_ny+C_nz\end{bmatrix}_{n\times 1}=\boldsymbol{M}_{n\times 3}\times\begin{bmatrix}x\\ y\\ z\end{bmatrix}\quad(1)$$

令

$$w_i=A_ix+B_iy+C_iz,i=1,2,\cdots,n\quad(2)$$

(3)MFFRW 算法。

MFFRW 算法将数据集中每张图片的 x,y,z 特征代入式 2 得到在 (A_i,B_i,C_i) 权值组合下由融合特征向量组成的数据集,将此数据按 2∶1 比例分成训练数据集和测试数据集,用训练数据集可训练得到一个分类模型,用测试数据集可得到此模型的分类正确率。如此,n 组随机权值就可以训练得到一个分类模型集合 $A=\{C_1,C_2,\cdots,C_n\}$ 和一个分类正确率集合 $B=\{T_1,T_2,\cdots,T_n\}$,将集合 A 中的数据降序排序后得到最优权值组合及其对应的 SVM 模型。

2 MFFRW 算法的并行设计及实现

2.1 Hadoop 平台

Hadoop[12] 平台是布式处理的软件框架,是一个被认可的用于处理海量各类型数据和应对复杂计算问题的平台。其中 HDFS 和 MapReduce 是 Hadoop 平台的两个核心设计。HDFS 是分布式文件系统,采用主/从模式体系结构,实现了对大规模数据集的流式访问。MapReduce 是一种并行编程模型,能够将计算任务和数据分配到 Hadoop 集群的各个节点上,它借助函数式编程方法,将计算分为 Map 和 Reduce 两个过程,每个过程的处理均以键值对的形式进行输入和输出,通过定义 Mapper 类和 Reduce 类实现一个键值对到另一个键值对的映射。

2.2 Hadoop 平台下的 MFFRW 算法

2.2.1 MFFRW 算法并行框架

在 Hadoop 平台上的 MFFRW 算法实现过程如图 2 所示。

整个流程分为 2 部分,第一部分完成图像 Hue 特征、纹理特征、PCA-Sift 特征的提取,第二部分完成特征融合并得到最优权值组合和对应的分类模型。在特征提取部分采用序列化文件的输入方式处理大量的图像小文件,降低 Hadoop 集群频繁启动 Map 任务的消耗,优化集群性能,提高图像特征提取效率。

在输出最优权值组合的过程中,通过重写 setup 方法、map 和 reduce 方法实现 MFFRW 算法的并行。同时,为了降低 Hadoop 系统的 I/O 消耗,只将模型路径输出到 HDFS 文件系统,最后通过 reduce 方法得到最优权值组合及其对应的分类模型路径。

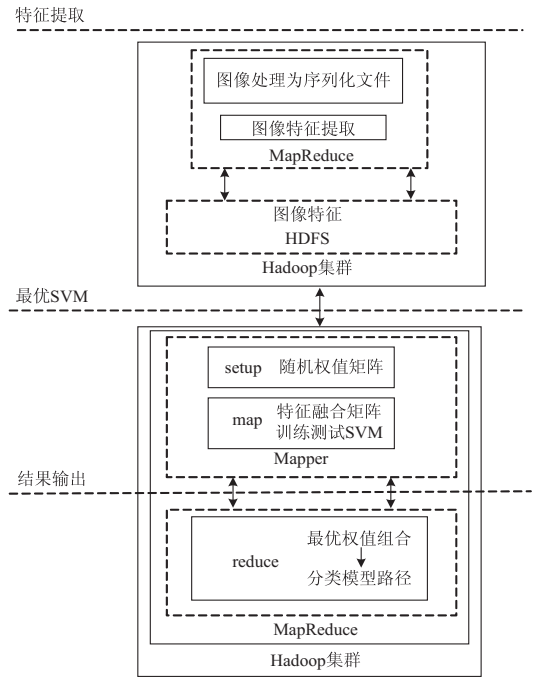

图 2 随机权值特征融合 SVM 图像分类总体框架

2.2.2 算法设计与实现

鉴于 OpenCV 是一个开源的、跨平台的计算机视觉库,实现了很多数字图像处理方面的算法,同时也提供了大量的 Java 接口[13]。因此,文中利用 OpenCV 的函数库实现图像特征提取和分类,并利用 PCA 算法对 Sift 特征降维,可以起到对特征向量去噪提纯的作用,提高匹配率[14]。Hadoop 平台下用 Java 语言编程实现了 MFFRW 算法。此外,由于 SVM 的各参数设置对分类结果影响很大,所以采用 K 折交叉验证算法得到径

向基函数的最佳参数,通常将 K 的值设为 10。

在 MapReduce 框架中,Map 任务和 Reduce 任务主要由 setup、map 或 reduce、cleanup 函数组成。其中 setup 函数只在任务开始时执行一次,map 和 reduce 函数循环执行多次直到全部数据处理完毕,cleanup 只在任务结束时执行一次。文中利用各函数的运行特点设计实现 MFFRW 算法,算法描述如下:

输入:n 个类别的图像训练和测试数据集 S_1, S_2,…,S_n。

特征提取算法:

Step1:将输入数据集中的小图像文件处理为序列化文件,保存至 HDFS 文件系统;

Step2:map 函数读取序列化文件并将图像统一处理为 150 * 200 大小,Step2 ~ 4 在 map 函数中实现;

Step3:获取每张图像的类别信息,并提取对应图像的 Hue 特征(x)、LBP 特征(y)、PCA - Sift 特征(z);

Step4:将 Step3 的类别信息和特征组成键值对形式<类别,特征值>,记为:cf_1, cf_2,…, cf_n,保存至 HDFS 文件系统。

MFFRW 算法:

Step1:在 Map 任务的 setup 函数中生成随机权值矩阵 $\boldsymbol{M}_{n\times 3}$,Step2 ~ 6 在 map 函数中实现;

Step2:将 $cf_i(i=1,2,\cdots,n)$ 拆分为 x, y, z 三个特征分量后代入式 2,得到融合后的特征向量集 WF;

Setp3:WF 按 2 ∶ 1 比例随机分为 WF_1 和 WF_2,选取 WF_1 作为训练样本, WF_2 作为测试样本;

Setp4:将 WF_1 的每个元素输入到 SVM,得到训练好的分类模型;

Setp5:将 WF_2 中的每个元素输入到训练好的分类模型,得到对应模型的分类正确率 Cr;

Setp6:反复执行 Step2 ~ 5,得到一个由 Cr 组成的集合 $\{Cr_1, Cr_2, \cdots, Cr_n\}$;

Setp7:reduce 函数中将 Setp6 集合中的数据进行排序,得到最大值和对应的权值组合,即最优权值组合。

输出:将最优权值组合和它对应的模型路径输出到 HDFS 文件系统。

3　实验结果及分析

3.1　实验环境和数据来源

利用 5 台计算机搭建 Hadoop 集群,1 台为 Master 节点,其余 4 台为 Slave 节点。所有节点计算机硬件配置都采用酷睿 i7 四核八线程 4.2 G 处理器,8 G 内存,4 T 硬盘空间;软件配置如下:操作系统为 64 位的 Ubuntu 14.04,Java 环境为 jdk1.7.0_79,Hadoop 为 Hadoop-2.5.1(64 位编译)的版本。

实验数据来源于 Corel1000 图像库中,每个类别的前 65 张图片作为训练集,后 35 张图片作为测试集,对于多个 SVM,训练和测试的数据总量情况如表 1 所示。

表 1　多个 SVM 训练、测试数据统计

SVM 个数	训练数据	测试数据	总计
50	32 500	17 500	50 000
80	52 000	28 000	80 000
100	65 000	35 000	100 000
120	78 000	42 000	120 000

3.2　图像分类及分析

3.2.1　图像分类正确率对比

为验证文中算法的效果,在相同条件下得到 MFFRW 算法与 1 ∶ 1 ∶ 1 融合、文献[6-7]的分类正确率对比结果,如表 2 所示。

表 2　正确率对比　　%

图像集	1 ∶ 1 ∶ 1 融合	文献[7]	文献[6]	文中算法
建筑	45	58	65	73
巴士	88	86	82	100
恐龙	94	92	91	100
大象	48	75	88	88
印第安人	61	68	72	76
花朵	61	77	89	94
食物	36	52	65	55
马	58	69	78	82
雪山	24	57	60	64
海边	36	60	47	52
平均正确率	56.5	69.4	73.7	78.1

从表 2 可以得知,文中算法的分类效果相对较好,巴士和恐龙的分类正确率达 100%,其中单个特征的分类正确率与 MFFRW 算法的对比效果如图 3 所示。

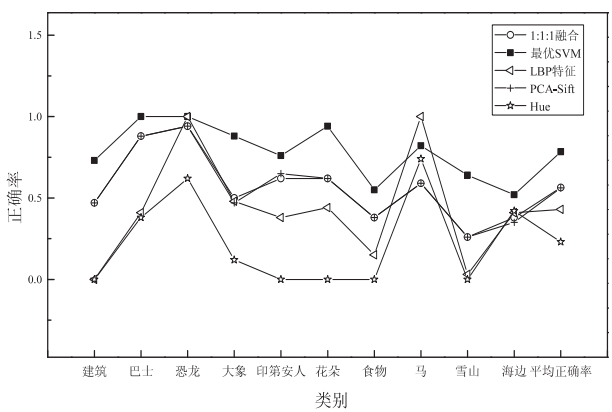

图 3　单个特征与最优 SVM 正确率对比

由图 3 可知,MFFRW 算法的分类效果比单个特

征更加优越。综上可知,通过文中算法确定的权值组合能更好地描述图像内容。

3.2.2 加速比

加速比[15]指同一任务在单节点环境下运行时间与多节点环境运行时间的比值,是衡量 Hadoop 平台下并行算法效率的一个重要指标。为验证文中算法在 Hadoop 平台下的性能,通过训练 50、80、100、120 个 SVM 进行了加速比实验,结果如图 4 所示。

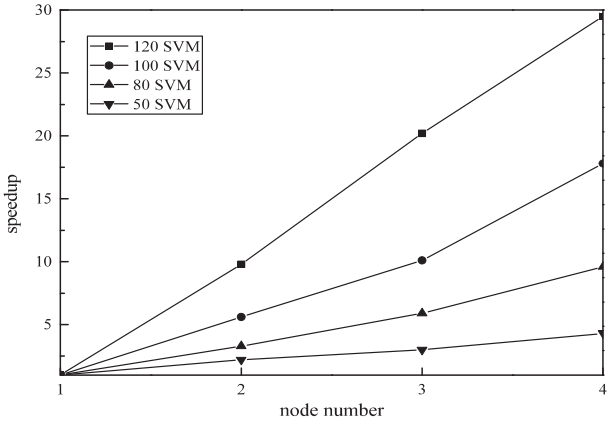

图 4 加速比对比

理想状态下,系统的加速比应随着节点计算机的增加而线性增长,但由于受通信开销、负载平衡的原因,实际上加速比并不能线性增长。从图 4 可以看到,在训练个数不多时,系统的加速比随着节点计算机的增多而增大,但增长幅度并不大,而随着训练个数的增多,系统加速比增长幅度会变大,当训练个数达到 120 时,系统的加速比几乎呈线性增长的趋势,这进一步说明了 Hadoop 集群在计算量大时更能体现其优越性。

4 结束语

对 Hadoop 平台下的 MFFRW 分类算法进行了深入的探讨和研究,并将 MapReduce 并行编程模型应用于文中算法以解决计算量大的问题,在保证算法正确率的前提下,有效提升了算法效率。实验结果表明,该算法分类正确率高、运行耗时少、不受主观因素影响,搭建的 Hadoop 集群能够充分利用各节点计算机的资源,相对于单节点计算机,系统获得了很好的加速比,充分体现了 Hadoop 集群分布式并行处理的强大运算能力。

随着大数据技术的广泛应用,将大数据技术应用于传统算法已成为新的研究热点。下一步的研究工作中,将扩展 Hadoop 集群的节点数,调节参数,提升分布式集群的性能;进一步改进图像特征提取算法。

参考文献:

[1] 吴云蔚,宁 芊. 基于 Hadoop 平台的分布式 SVM 参数寻优[J]. 计算机工程与科学,2017,39(6):1042-1047.

[2] AZHAR R, TUWOHINGIDE D, KAMUDI D, et al. Batik image classification using SIFT feature extraction, bag of features and support vector machine[J]. Procedia Computer Science, 2015, 72: 24-30.

[3] CAMLICA Z, TIZHOOSH H R, KHALVATI F. Medical image classification via SVM using LBP features from saliency-based folded data[C]//IEEE 14th international conference on machine learning and applications. Miami, FL, USA: IEEE, 2015: 128-132.

[4] 陈苏婷,王 慧. 多尺度多特征融合的高分辨率遥感影像分类[J]. 量子电子学报,2016,33(4):420-426.

[5] 李 静,贾利民. 数据融合综述[J]. 交通标准化,2007(9):192-195.

[6] 胡湘萍. 基于多核学习的多特征融合图像分类研究[J]. 计算机工程与应用,2016,52(5):194-198.

[7] 许庆勇,江顺亮,黄 伟,等. 基于多特征融合的深度置信网络图像分类算法[J]. 计算机工程,2015,41(11):245-252.

[8] 张春森,郑艺惟,黄小兵,等. 高光谱影像光谱-空间多特征加权概率融合分类[J]. 测绘学报,2015,44(8):909-918.

[9] 李玉峰,郑德权,赵铁军. 基于 SVM 和多特征融合的图像分类[C]//第四届全国信息检索与内容安全学术会议论文集(上). 北京:中国学术期刊电子出版社,2008:630-635.

[10] 袁广林,薛模根,韩裕生,等. 基于自适应多特征融合的 mean shift 目标跟踪[J]. 计算机研究与发展,2010,47(9):1663-1671.

[11] 刘 爽. 多特征融合图像检索方法及其应用研究[D]. 哈尔滨:哈尔滨理工大学,2016.

[12] WHITE T. Hadoop: the definitive guide[M]. 3rd ed. California, USA: O'Reilly Media, 2012.

[13] ALJASEM D K, HEENEY M, GRITTI A P, et al. On-the-fly image classification to help blind people[C]//12th international conference on intelligent environments. London, UK: IEEE, 2016: 155-158.

[14] 李 钦,游 雄,李 科,等. PCA-SIFT 特征匹配算法研究[J]. 测绘工程,2016,25(4):19-24.

[15] 韩 伟,张学庆,陈 旸. 基于 MapReduce 的图像分类方法[J]. 计算机应用,2014,34(6):1600-1603.