文章编号：2095-4980(2014)05-0740-06

云计算下基于贝叶斯网络的多传感器目标识别

张　明，李　波，刘学全，高晓光

(西北工业大学 电子信息学院，陕西 西安 710129)

摘　要：基于云计算可实现分布式并行程序海量数据处理的特点，提出将多传感器目标识别融合处理部署在云计算 Hadoop 平台上，并将其运行在多个节点组成的计算机集群上。根据目标识别原理建立贝叶斯网络结构，对目标识别预处理得到的数据进行融合计算，推理目标类型，并对不同情况下的 Hadoop 集群效率进行分析比对。实验结果证明了将目标识别融合处理部署在云计算平台上可有效提升运算效率。

关键词：目标识别；云计算；Hadoop 平台；贝叶斯网络；多传感器；信息融合

中图分类号：TN911.73；TP181　　**文献标识码：**A　　**doi：**10.11805/TKYDA201405.0740

Multi-sensor target recognition based on Bayesian classifier in cloud computing environment

ZHANG Ming，LI Bo，LIU Xue-quan, GAO Xiao-guang

(School of Electronic & Information，Northwest Polytechnic University，Xi'an Shaanxi 710129，China)

Abstract：Based on the cloud computing characteristics of distributed parallel processing of massive data, multi-sensor target recognition on the Hadoop platform is put forward, which consists of a plurality of nodes running on computer clusters. Bayesian network structure is established according to target recognition theory. The data obtained by the pre-process of object recognition are fusion-calculated to deduce the target type. The efficiencies of Hadoop clusters under different circumstances are compared. Experimental results show that running multi-sensor information fusion target recognition on cloud computing platform is intuitive and effective, which successfully improves the operational efficiency.

Key words: target recognition; cloud computing; Hadoop; Bayesian network; multi-sensor; information fusion

随着微电子技术、信号检测与处理技术、计算机技术、网络通信技术以及控制技术的飞速发展，各种面向复杂应用背景的多传感器目标识别系统大量涌现，包括对多种目标的探测、识别、跟踪等[1]；然而现代战争空地作战一体化的特点及作战双方采取的各种电子对抗技术，使战场环境日益复杂恶劣，单一传感器所获得的数据不精确，不完整，不可靠。利用多传感器实现目标识别，更大程度地收集和处理目标和环境的信息，从而提高识别系统的准确性和可靠性。在这个过程中，随着传感器传输信息量的激增，冗余和互补信息的融合，同样也引入了“信息冲突”这个新问题，原因在于各传感器存在大量无用信息以及信息的不确定性和模糊性。云计算平台提供一种能够对海量数据进行分布式处理的平台服务，其具有可靠、高效、易扩展和高容错性的特点。将多传感器目标识别部署到云计算平台上，既能提升目标识别效率，又能有效地解决多传感器信息中存在的不确定性和模糊性等问题。本文选取朴素贝叶斯网络作为目标识别算法，采用Hadoop云计算平台作为目标识别数据处理平台，通过对比不同情况下Hadoop集群的数据处理效率，验证目标识别数据处理部署在云计算平台上的可行性及效率提升。

1　目标识别分析与相关算法

1.1 空中目标识别分析

敌方空袭兵器突击目标，其行为特征具有明显的规律。不同的空袭兵器的行为特征在某些因素上反映突出，

收稿日期：2013-11-05；**修回日期：**2013-11-24

基金项目：航天技术支撑基金资助项目(2013-HT-XGD)；西北工业大学基础研究基金资助项目(JC201144)

另一些因素比较模糊。因此，在目标识别中，可选取对类型反映突出的因素作为主要识别因素，依据平时收集的敌人资料，再加上已有的经验和空袭样式发展的预测来估计，具有下列指标集：

1) 飞行高度(Height)：可分为大于 7 000 m，500 m~7 000 m，小于 500 m 三档；

2) 飞行速度(Speed)：战术弹道导弹(Tactical Ballistic Missile，TBM)的再入速度一般在 1 800 m/s~2 200 m/s；武装直升机的速度小于 100 m/s；一般作战飞行器的突防速度为 200 m/s~400 m/s；可分为大于 1 200 m/s,600 m/s~1 200 m/s,200 m/s ~600 m/s，小于 200 m/s 四档；

3) 发现距离(Dis)：分为大于 150 m,150 m~100 m,100 m~50 m，小于 50 m；

4) 航线特征(Character)：分为等高平直飞行、俯冲、下滑。

由于目标飞行中，常带有随机性和模糊性，推理规则和推理过程也带有模糊性，依据技术资料和专家经验，一般有如下的目标类型：a) 若飞行高度较高，且发现距离较远，速度又较慢，则目标可能为轰炸机；b) 若飞行高度中高，且发现距离较远，速度又中速，则目标可能为歼击机；c) 若发现距离较近，且速度较快，则目标可能为空地导弹或反辐射导弹；d) 若飞行高度较高，且发现距离较近，速度又很快，则目标可能为 TBM；e) 若飞行高度较低，且速度很慢，则目标可能为武装直升机[2]。

1.2 目标识别结构与相关算法

目标识别融合有3种处理结构，即融合出现在原始数据级(特征提取之前)，特征矢量级(属性说明之前)，判决级(各传感器给出独立属性判决之后)。

数据级融合法。在这种方法中，匹配的传感器数据直接融合，然后对融合的数据进行特征提取和属性说明。实现数据级融合的传感器必须是相同的(如若干IR (International Rectifier)传感器)或匹配的(如1个IR传感器和可见光成像传感器)，在原始数据上实现关联保证了与同一目标或实体有关的数据进行融合。传感器的原始数据融合后，识别的处理等价于对单个传感器的处理。对于成像传感器，数据级融合经常称为像素级融合。数据级融合所达到的精确度依赖于可得到的物理模型的精确度，这类技术包括像Kalman滤波那样的仿真和估计法。

特征级融合法。在这种方法中，每个传感器观测目标，并对各传感器的观测进行特征提取，产生特征矢量，然后融合这些特征矢量，并作出基于联合特征矢量的属性说明。实现属性说明处理的技术包括聚类分析、神经网络、模板及基于知识的技术。在这类方法中，必须通过关联处理把特征矢量分成有意义的组。定位信息对这个关联处理可能是有意义的，因为实际中特征矢量可能是非常广的量(如傅里叶参数、时域特征等)。

判决级融合法。首先，每个传感器各自进行变换，独立地产生属性说明，而后融合各传感器的属性说明，进而形成一个联合的属性说明。其融合技术包括表决法、Bayesian推理、Dempster–Shafer法、广义证据处理理论及其他方法[3]。本文应用此方法。

2　目标识别与云计算

多传感器目标识别技术是利用计算机技术对按时序获得的若干传感器的观测信息,以及数据库和知识库的信息，在一定准则下加以自动汇集、相关、分析、综合为一种表示形式，以完成所需要的估计和决策任务所进行的信息处理过程。目前的目标识别数据融合技术不仅涵盖了声、光、电等物理层数据的处理，而且涉及到了数据库、网页、视频、资讯、自然语言等较高层次的信息整合，因此数据融合也称为信息融合[4]。由于多传感器目标识别信息的冗余性、互补性、时效性和低代价，多传感器系统克服了单一传感器的局限。然而在这个过程中，冗余和互补信息的融合，同样也引入了“信息冲突”这个新问题，原因在于各传感器存在大量无用信息以及信息的不确定性和模糊性[5]。

云计算Hadoop平台可以对海量日志、文档或数据进行预处理，依据需求提取有效字段并通过质量代码检测读取的数据是否可疑或错误，降低目标识别信息融合过程中多传感器的无效信息；其通过并行处理方式将海量数据部署在分布式集群中进行数据处理，而且它的设计思想是以流式数据访问存储超大文件，运行在商用硬件集群上，具有高冗余性、容错性和效率，这些特点为目标识别数据融合提供了有效保障。

3　目标识别算法的云计算实现

3.1 云计算开源平台——Hadoop

Hadoop是Apache基金会下的一款开源软件，它实现了包括分布式文件系统HDFS(Hadoop Distributed

Filesystem)和MapReduce[6]框架在内的云计算软件平台的基础架构，整合了包括数据库、云计算管理、数据仓储等一系列平台，已成为工业界和学术界进行云计算应用和研究的标准平台[7]。

1) HDFS 是 Hadoop 的旗舰级分布式文件系统，实际上 Hadoop 是一个综合性文件系统抽象，因此 Hadoop 也会继承其他文件系统。其主要包含数据块、NameNode 和 DataNode、客户端 3 个概念以及块存储和冗余复制技术。

2) MapReduce 是一个用于数据处理的分布式编程模型，它将任务调度、容错机制、空间局部最优化和节点间的负载均衡在模型内部实现，用户只需要提供 2 个接口函数：map(映射)和 reduce(规约)。map 和 reduce 之间以键值对的形式接收和传输数据，map 是映射过程，将 split 中的数据以键值对的形式读入，并输出新的键值对。不同 map 产生的相同的键的值会以数组形式传给 reduce 进行规约，reduce 接收到键值对并对其处理后得到最终的结果[8]。即：

Map:<key1,value1> → <key2,list(values)>

Reduce:<key2, list(values)> → output

依据Hadoop并行原理以及贝叶斯网络识别精确度高，识别速度快，区分度好的特点，本文选择朴素贝叶斯网络为目标识别算法。

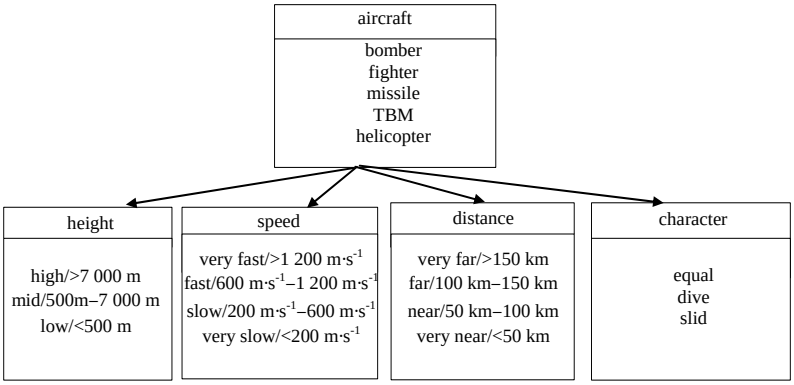

Fig.1 Naive Bayesian network of target recognition
图1 目标识别的朴素贝叶斯网络

3.2 朴素贝叶斯网络的建立

对于贝叶斯网络结构模型的建立，首先要确定网络中的假定变量、观测变量和中间变量。根据对空中目标识别的分析，建立如下的网络模型(见图1)。假定变量为识别的目标类型，其取值为所有可能的目标平台类型，用根节点表示。假定所有待识别的目标平台类型为轰炸机、歼击机、空地导弹、弹道导弹与武装直升机；观测变量为各类传感器观测到的目标平台的运动特征等，比如雷达可以获得目标的距离、速度和形态等，用子节点来表示。

假设目标数据集中所有样本都可以用一个N维特征向量 $\boldsymbol{X}_j=\{x_1,x_2,\cdots,x_n\}$ 表示，由贝叶斯公式计算得到的新样本的后验概率用 $P(C_i\mid\boldsymbol{X}_j)$ 表示：

$$P(C_i\mid X_j)=\frac{P(X_j\mid C_i)\times P(C_i)}{P(X_j)} \tag{1}$$

朴素贝叶斯的基本假设是属性之间相互独立且顺序无关，故：

$$P(\boldsymbol{X}_j\mid C_i)=P(\{x_1,x_2,\cdots,x_n\}\mid C_i)=\prod_{k=1}^{n}P(X_k\mid C_i) \tag{2}$$

由式(1)、式(2)得：

$$P(C_i\mid\boldsymbol{X}_j)=\frac{\prod_{k=1}^{n}P\left(\boldsymbol{X}_k|C_i\right)\times P\left(C_i\right)}{P\left(\boldsymbol{X}_j\right)} \tag{3}$$

对于测试集中某个样本 $\boldsymbol{X}_j$，分类器将选出使其后验概率达到最大的那个类别。而 $P(\boldsymbol{X}_j)$ 对于每个类别都相同，不会影响最终结果，可以不计算，公式为：

$$C\left(\boldsymbol{X}_j\right)=\operatorname{argmax}_{C_i}\left(P(C_i\mid\boldsymbol{X}_j)\right)=\operatorname{argmax}_{C_i}\left(\prod_{k=1}^{n}P\left(\boldsymbol{X}_k|C_i\right)\times P\left(C_i\right)\right) \tag{4}$$

$P(X_k\mid C_i)$ 表示属性 X_k 属于类别 C_i 的概率，$P(C_i)$ 表示类别 C_i 的条件概率，由训练集中的样本进行估算。

$$P(C_i)=\frac{D_i}{D} \tag{5}$$

$$P(X_k\mid C_i)=\frac{N_{kc}+1}{N_c+\left|V\right|} \tag{6}$$

式中：D_i 表示类 C_i 中样本数；D 表示训练集总样本数；N_{kc} 是属性 X_k 在类 C_i 中出现的样本数；N_c 是类中样本总数；$\left|V\right|$ 是总的类别数。$P(X_k\mid C_i)$ 采用了拉普拉斯修正，以防止 $P(X_k\mid C_i)=0$ (Laplace=1)。

3.3 算法实现

本文目标识别中朴素贝叶斯算法可以分为2部分：目标数据预处理(从传感器数据中提取有效数据)；训练过程(进行参数学习)。

其中算法的并行,采用了单一MapReduce作业,通过Mapper完成对目标数据预处理，并将Mapper处理好的数据进行二次排序，即将相同类别目标进行分区然后依据不同属性进行分组，整组传入Reducer中进行总数计算并求解先验概率，流程如图2所示。

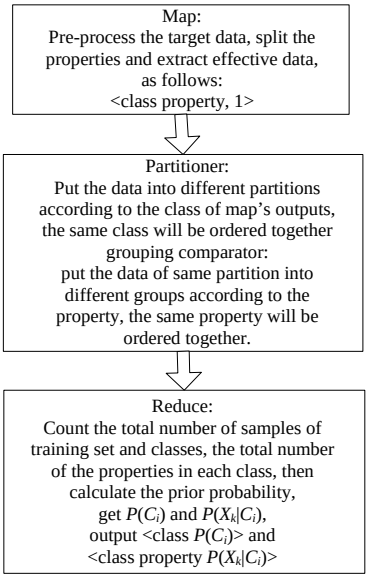

Fig.2 Process of Bayesian network parallel implementation
图2 贝叶斯网络并行实现流程

4 实验分析

4.1 仿真硬件环境

仿真硬件环境见表 1。

表1 仿真硬件环境
Table1 Hardware environment of simulation

environment	specific configuration
CPU	i5-2320 CPU @ 3.00 GHz
memory	DDR3-1333 1 024 MB
hard drive space	200 GB
network	100 MB Ethernet
operating system	Microsoft Windows XP Professional edition 2002 Service Pack 3
Hadoop edition	Hadoop-0.20.2

4.2 不同数据集在单机与集群中运行时间比较

本实验选取了6组大小不同的训练集(见表2)，并以相同的程序进行测试。其中训练集大小分别为：48.1 MB,80 MB,348 MB,822 MB,1.5 GB和2 GB。通过比较单机和4个节点组成的集群对不同大小数据集的处理情况，来比较并行带来的速度提升。

从图3中可得出结论：4个节点的集群随着数据集的增加，时间变化相对较缓慢；而单机在处理数据时，会随着数据集的增长呈现出时间大幅增长并超过相同数据集下集群所用时间的情况。

原因分析：HDFS是以数据块(默认大小为64 MB)存储的，每个文件都会按照数据块大小来进行分割处理，每块分配给一个map任务。当数据集较小时，任务处理时间短，集群节点之间的管理和通信占据较大时间开销，造成单机的处理时间短于集群的处理时间。而随着数据的增加，任务的处理时间增加，单机处理任务的时间远远长于并行处理任务的时间，这样多节点间管理和通信开销相对较小，使得多节点的任务处理效率高于单节点。

表2 不同数据集在单机与集群中运行时间
Table2 Different data process time at 1or 4 nodes

	48.1 MB	80 MB	348 MB	822 MB	1.5 GB	2 GB
single/s	79	91	346	541	1 007	1 272
4 nodes/s	97	117	287	404	690	990

表3 不同节点数下数据处理时间
Table3 Processing time at different number of nodes

nodes	1	2	3	4	5
time/s	541	321	274	354	463

4.3 不同数量节点时效率比较

本实验选取802 MB的目标识别数据训练集分别在1节点、2节点、3节点、4节点集群中进行测试，并通过相同并行算法程序对该训练集进行处理，进而比较不同节点数对相同数据进行处理的效率差别(见表3)。

由图4可得出结论：在处理大数据集时，多节点的效率高于单节点，而在数据一定时，存在一个节点数其效率最高。

原因分析：在处理大规模数据时单一节点的数据处理性能有限，这时将运算量平均分布到多节点上并行运算，能够有效地解决单一节点的效率低下问题。而在集群节点对效率提升过程中因存在一个临界值点，其效率

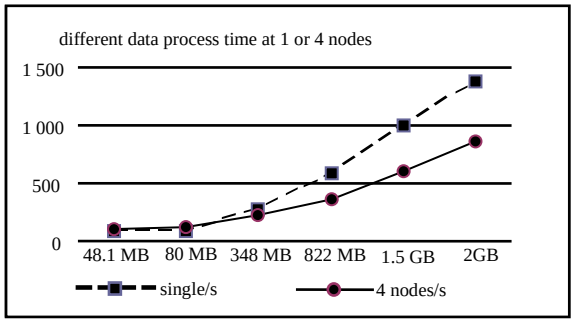

Fig.3 Different data process time at 1 or 4 nodes
图 3 单机与 4 节点处理不同数据的效率对比

最高，节点数低于临界值则节点效率低，高于临界值则造成节点运算资源的浪费并且增加了节点之间的管理和通信开销，从而降低整体效率。

4.4 目标识别可行性分析

从经过预处理的目标识别样本集中随机抽取一组歼击机数据进行推理验证，选取数据如下：目标在10 000 m的高空飞行，飞行物大约以279 m/s的速度飞来，雷达探测到的目标距离为150 km，飞行物正在平飞，依据贝叶斯分类算法计算出后验概率，经贝叶斯网络的融合识别结果见图5。

以上经过推理得出的结果与数据中类别相符，可行性得到了验证。

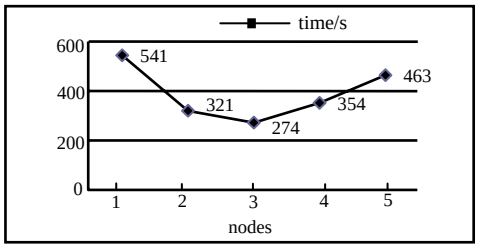

Fig.4 Processing time at different number of nodes
图 4 不同节点下数据处理时间对比

5 结论

本文初步探讨了云计算在目标识别领域应用的可行性，并以朴素贝叶斯网络为算法的目标识别数据处理并行化为例，将其部署在云计算Hadoop平台上，然后对不同情况下的Hadoop集群的数据处理效率进行比对，并验证了目标识别可行性。仿真结果表明：将目标识别问题部署在云计算平台上可行；其次Hadoop适合处理大数据，随着节点数的增加，贝叶斯算法在Hadoop上表现出较高的效率，而在处理小数据时，节点数的增加反而会使其效率降低；最后，在处理相同大小的数据集时，多节点的效率高于单节点，并在某特定节点数时效率最高。为了提高任务处理效率，而无限度地增大集群规模，会增加管理和节点之间的通信成本，甚至可能导致死锁，严重影响数据处理效率。选择合适的集群节点数对于提高数据处理效率具有一定影响。

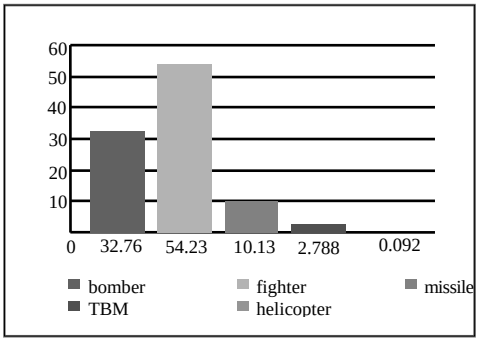

Fig.5 Recognition results of Bayesian network
图 5 贝叶斯网络融合识别结果

参考文献：

[1] 潘泉,于昕,程咏梅,等. 信息融合理论的基本方法与进展[J]. 自动化学报, 2003,29(4):599–601. (PAN Quan,YU Xin, CHENG Yong-mei,et al. Essential methods and progress of information fusion theory[J]. Acta Automatica Sinica, 2003, 29(4):599–601.)

[2] 王安丽,史志富,张安. 基于熵的空中目标识别模型及应用[J]. 火力与指挥控制, 2005,30(2):110–112. (WANG An-li, SHI Zhi-fu,ZHANG An. Model and application to target recognition of based on entropy[J]. Fire Control and Command Control, 2005,30(2):100–112.)

[3] 马平,吕锋,杜海莲,等. 多传感器信息融合基本原理及应用[J]. 控制工程, 2006,13(1):48–51. (MA Ping,LV Feng,DU Hai-lian,et al. Theory and application of multi-sensor information fusion[J]. Control Engineering of China, 2006,13(1):48–51.)

[4] 孙辉,赵峰,张峰云. 多传感器信息融合技术及其应用[J]. 海洋测绘, 2009,29(5):77–81. (SUN Hui,ZHAO Feng ZHANG Feng-yun. Multi-sensor information fusion technology and its application[J]. Mechanical Management and Development, 2009,29(5):77–81.)

[5] 杨莘元,蒲书缙,马惠珠. 多传感器目标识别的优化融合[J]. 宇航学报, 2005,26(1):47–51. (YANG Shen-yuan,PU Shu-jin,MA Hui-zhu. Optimized fusion in multi-sensor target recognition[J]. Journal of Astronautics, 2005,26(1):47–51.)

[6] 马文芳. Hadoop:云中起舞的小象[J]. 计算机报, 2010(43):28–30. (MA Wen-fang. Hadoop:the calf elephant dancing in the cloud[J]. Chinese Journal of Computers, 2010(43):28–30.)

[7] Dean J,Ghemawat S. Distributed programming with Mapreduce[A]// Oram A,Wilson G. Beautiful Code[M]. Sebastopol:O'Reilly Media, Inc, 2007: 371–384.

[8] Tom White,周傲英,曾大聃. Hadoop权威指南[M]. 北京:清华大学出版社, 2011. (Tom White,ZHOU Ao-ying,ZENG Da-dan. Hadoop:The Definitive Guide[M]. Beijing:Tsinghua University Press, 2011.)

(下转第 766 页)