Computer Science and Application 计算机科学与应用, 2015, 5(9), 336-342
Published Online September 2015 in Hans. http://www.hanspub.org/journal/csa
http://dx.doi.org/10.12677/csa.2015.59042

Hans汉斯

Security Technology of the Cloud Big Data Based on Deep Learning

Tiankai Sun[1,2*], Rong Bao[1], Daihong Jiang[1], Kui Wang[1]

[1]School of Information and Electrical Engineering, Xuzhou Institute of Technology, Xuzhou Jiangsu
[2]Faculty of Electronic Information and Electrical Engineering, Dalian University of Technology, Dalian Liaoning
Email: *strongtiankai@163.com

Received: Oct. 5th, 2015; accepted: Oct. 23rd, 2015; published: Oct. 29th, 2015

Abstract

The cloud big data is the basis of the data analysis. The security and accuracy of the big data is essential to the result of data analysis. By combining Hadoop's big data processing technology and digital watermarking technology, a classification with DBN as a smart strategy is proposed. The multilayer has been trained and adjusted by this scheme. The mass of data can be calculated and the distributed data can also be obtained which is the basis of the judgment of data tampering. The experiments show that the combination of Hadoop and AI is an effective method to the massive data security.

Keywords

DBN, Data Analysis, Hadoop, Intelligent Classification

基于深度学习的云端大数据安全防护技术

孙天凯[1,2*]，鲍　蓉[1]，姜代红[1]，王　奎[1]

[1]徐州工程学院信电工程学院，江苏　徐州
[2]大连理工大学电信学部，辽宁　大连
Email: *strongtiankai@163.com

收稿日期：2015年10月5日；录用日期：2015年10月23日；发布日期：2015年10月29日

*通讯作者。

文章引用: 孙天凯, 鲍蓉, 姜代红, 王奎. 基于深度学习的云端大数据安全防护技术[J]. 计算机科学与应用, 2015, 5(9): 336-342. http://dx.doi.org/10.12677/csa.2015.59042

摘　要

云端海量大数据是数据分析的基础，数据本身的安全性和准确性，对数据分析的结果有重要影响。针对云端大数据的特性，融合Hadoop的海量大数据处理以及数字水印相关技术，提出了一种以深度信念网络(DBN)作为智能分类的机制，通过对数据进行多层的训练和调整，对云端海量数据进行计算，得到其分布式表示，进而获取数据的篡改和判断的依据。实验表明，Hadoop和AI的结合，很好的实现了云端海量大数据的安全防护。

关键词

DBN，数据分析，Hadoop，智能分类

1. 引言

伴随着云端大数据[1]时代的到来，传统的关系型数据处理技术已无力处理海量云端大数据。当前，已有的智能化设备仍然不能像人脑一样进行智能化的学习和干预事务。数据就是命脉，如何以最快的速度响应处理这些数据，如何保障这些海量数据的安全，成为当前研究的一大热点问题。

数据的正确性和完整性是海量数据分析的基础。传统方式下的数据安全保护技术，一般是进行数据隐藏或者信息加密，在数据的处理量上，往往也是以少量数据的处理为主，仅仅采用以往技术，很难解决当下云端海量大数据的安全防护问题。与此同时，在智能化防护性上，例如大数据的智能化分类、被篡改数据的智能化识别、智能化学习以及智能化定位等，传统的数据保护技术很难解决。深度学习(Deep Learning) [2]相关技术的发展，提供了一种新的处理云端大数据安全防护问题的思路。最近几年来，国内外许多学者专家对 Deep Learning 进行了深入的研究和探索[3]。与此同时，Hadoop [4] [5]的快速发展为海量数据的安全保护又提供了一个高效的备选方案。但这两种新的技术，并没有进行相互的结合，只是在各自的领域有所发展。

融合 Hadoop 的大数据处理技术以及人工智能中的深度学习技术，对疑是遭受篡改的数据进行智能化的识别和分类处理，与此同时在进行海量处理过程中，使用分布式的技术方案，实现了快速、准确定位篡改数据，快速恢复被篡改的数据的目标。

2. 智能分类模型

2.1. 受限玻尔兹曼机

Deep Learning 是一个多层的神经网络，是模拟人脑进行分析和学习。深度学习的模型主要是含多隐层的多层感知器，通过将低层的特征，进行抽象组合处理，得到抽象表示的高层数据。这样的逐层抽象和认知的过程，形成分布式表示的数据。

受限玻尔兹曼机[6] (RBM)如图 1 所示，是一种马尔可夫随机场。它由可视层 v 和隐含层 h 构成，并且可视层和隐含层都是条件独立的，即：

$$p(h|v)=\prod_i p(h_i|v)$$
$$p(v|h)=\prod_j p(v_j|h) \quad (1)$$

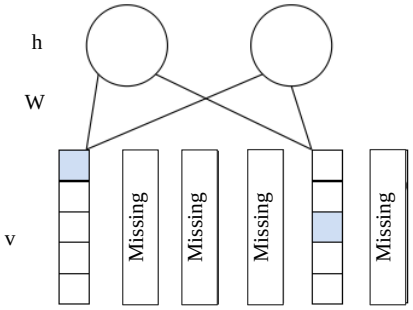

Figure 1. Restrict Boltzmann Machines, where v is visible unit, h is hidden unit, and W is weight matrix
图 1. 受限玻尔兹曼机(其中 v 为可见单元,h 为隐单元,W 为连接权重)

输入 v，通过公式(1)中的 $P(h|v)$ 以及 $p(v|h)$ 可求得隐藏层 h，以及可视层 v_1，通过参数的修正，以期达到从隐藏层得到的可视层 v_1 与原来的可视层 v 尽可能的保持一致，由此得到的隐藏层是原有可视层的另外一种表达。为了训练完成该神经网络，得到可视层节点和隐节点间的权重 W 和偏离量 b、c，需要引入能量模型。RBM 的能量模型定义为：

$$E(v,h)=-\sum_{i=1}^{n}\sum_{j=1}^{m}w_{ij}h_i v_j-\sum_{j=1}^{m}b_j v_j-\sum_{i=1}^{n}c_i h_i \tag{2}$$

当系统的总能量最小时，系统越稳定，并且极大似然取得的值最大，这样可以通过极大似然估计来求解系统参数。首先得求得可视节点和隐含节点的条件概率：

$$\begin{aligned} p(v_i=1|h,\theta)&=\frac{1}{1+\exp\left(-\sum_j w_{ij}h_j-b_i\right)}\\ p(h_j=1|v,\theta)&=\frac{1}{1+\exp\left(-\sum_i w_{ij}v_i-c_j\right)} \end{aligned} \tag{3}$$

其中 $\theta=\{w,b,c\}$，为参数集合，通过 Gibbs 采样，可以求得参数 w、b 和 c 的值。

2.2. 深度信念网络

通过 *RBM* 可以组成深度信念网络(DBN) [7]，DBN 模型与传统的判别模型相对，是一种概率生成模型，通过对 $P(Observation|Label)$ 和 $P(Label|Observation)$ 都做评估，而建立一个标签和观察数据之间的联合分布，而判别模型只评估后者，也就是 $P(Label|Observation)$。DBN 的结构[7]如图 2 所示。

深度信念网络其工作过程分为两个阶段，第一阶段为逐层构建单层神经元，这样每次都是训练一个单层网络，第二阶段为使用 Wake-Sleep 算法进行调优。训练过程如下：

1) 训练第一个受限玻尔兹曼机；

2) 设定第一个 RBM 的偏移量 b_1，c_1 的值以及权重 w_1，将隐含层神经元的状态值，作为临近第二个 *RBM* 的输入；

3) 第二个 RBM 被训练完之后，把第二个 RBM 相关信息堆叠到第一个 RBM 之上；

4) 重复之前步骤多次，堆叠多个 RBM；

5) 如果训练集中有标签数据，在对顶层进行训练时，把无标签数据和分类标签一起进行训练；

调优过程如下：

1) 除顶层之外，其它层间的权重为双向的，即，同时具有生成权重以及认知权重；

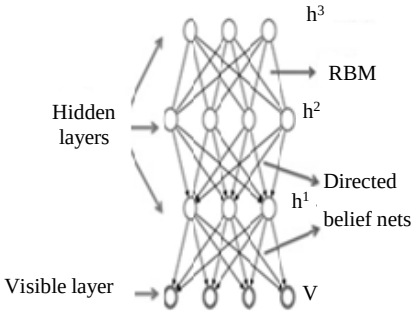

Figure 2. The structure of DBN
图 2. DBN 结构图

2) Wake 阶段：借用认知权重(向上的权重)和外界的特征来产生每一层的结点状态(抽象表示)，同时使用梯度下降的相关方法来修改层间的生成权重(下行权重)。

3) Sleep 阶段：通过向下权重值和顶层表示的状态(醒时学得的概念)，产生底层的状态，层间向上的权重值同时被修改。

2.3. 智能分类策略

对于云端海量大数据的篡改判断以及准确定位，需要考虑判断和定位的准确性和效率问题，而深度信念网络(DBN)在顶层，可以通过带标签数据，使用 BP 算法对判别性能做调整，同时，被附加到顶层的标签数据，会被推广到联想记忆中，并且，通过多层自下而上的受限玻尔兹曼机的训练，学习到的识别权值将获得一个网络的分类面，结合联合记忆内容，可以准确判断数据的篡改情况，从而进一步定位目标。

具体的智能分类策略如下：

1) 使用数字水印技术为需要保护的数据生成原始标签，供深度学习训练；

2) 使用 DBN 训练海量数据，获得各层的特征表示、权重矩阵以及偏移量等；

3) 将需要判断篡改的海量数据经 Map-Reduce [8]预处理，得到 DBN 的输入神经元；

4) 对输入的神经元数据，逐层进行 RBM 认知训练，同时做好向下的生成训练；在最后一层，结合第一步获得的标签数据以及之前训练获得的联想记忆，进行分类；

5) 对分类结果作分析处理，得到篡改数据的二维坐标，从而可以准确定位篡改数据的位置。

3. 海量数据处理技术

3.1. HDFS 分布式文件系统

在海量数据的处理方面，使用了 Apache 的开源框架 Hadoop，利用 Hadoop 的分布式存储和并行计算框架，处理海量级别数据的同时，保证了处理的速度以及高可用。

就需要保护的海量数据而言，在存放时，如果将数据存储在某一单个设备上，一旦设备出故障，丢失数据的可能性非常大，而且单个存储设备内，往往很难容纳这么大的数据量。针对上述问题，Hadoop 的 HDFS [4] [5] [9]提供了一种较好的解决方案。HDFS 主要由 DataNode、NameNode 以及 Client 组成，DataNode 是文件存储的基本单位，它将 Block 数据信息存储在本地文件系统中，以此保存了 Block 的 Meta-data 数据信息，与此同时周期性地将所有存在的 Block 数据信息发送给 NameNode。NameNode 作为分布式文件系统中的管理者，它的职责是管理存储块的复制、文件系统的命名空间以及集群配置信息等。Client 是一种应用程序主要用来获取分布式文件系统的文件。

3.2. Map-Reduce 分布式并行计算框架

海量数据被分割，有序存储到各个节点中时，各个节点上的 CPU、内存等资源，都可以被充分的使用。与此同时，调用 hadoop 的并行计算框架，各个节点上的资源由系统统一分配、调度，在各自的节点上完成数据的计算，新的任务到来时，如果该节点计算繁忙，将不会被分配任务，同时任务被分配给那些计算空闲的节点，以此实现负载均衡。

该框架的思想是：将 JobTracker 的两个主要功能进行分离，分离为单独的组件：任务调度/监控和资源管理。资源管理器具有全局管理所有的应用程序的计算以及资源配置的功能，而每一个 Application Master 负责相应的协调和调度工作。每一台机器的节点以及 Resource Manger 管理服务器，以此实现管理用户在那台机器上的进程的工作。

Map 过程：按键值对(*key*/*value*)将分割的数据进行映射处理，其中，每条记录的主键值通过 *hash* 计算求得 *key* 的值，需要保护字段的特征值即为 *value* 值。

Reduce 过程：依据等值的 key 值，将每个 value 的结果串结成一条虚拟链，再对形成的虚拟链进一步处理，得到具有新特征值的 value。

详细的 Map-Reduce 过程[4] [5] [9]如下图 3 所示。

4. 仿真模拟

为了验证算法的有效性及相关性能，我们选择了一张含有 2000 余万条记录、49 个列的大表进行模拟实验，以一个双节点的 linux 集群为实验平台，以 Java 为开发语言，Hadoop 的版本为 2.2.0，开发 IDE 为 eclipse，集群配置 JDK 的版本为 1.6。

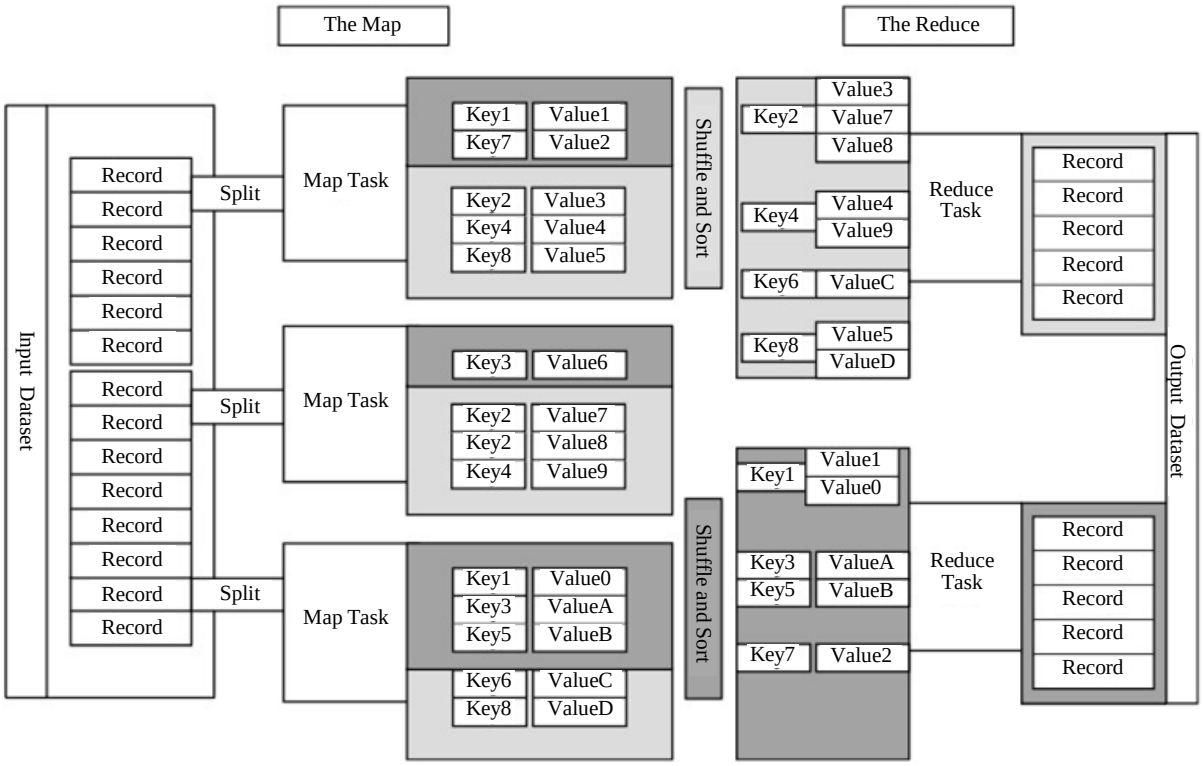

Figure 3. The structure of Map-Reduce
图 3. Map-Reduce 结构

实验结果和分析

在使用 DBN 网络进行模拟实验时，由于 RBM 的特点，所有节点使用 0、1 进行表示，1 表示节点被激活，0 表示节点没被激活，在训练的过程中，可视节点和隐含节点多次被重构，并且如果重构的结果和实际不太一样，将进行调整，以达到尽可能的相同，即数据的另外一种特征表示。

同时，在实验时，使用了 RBM 和 BP 两种算法进行了对比实验，从而尽可能的得到两种算法的特点和效率。

下图 4 使用的是 BP 算法进行模拟训练，图 5 使用的是 RBM 进行训练，其中红色的线表示当前错误率的变化趋势，蓝色的线表示错误率的改善趋势。

表 1 给出了 BP 算法和 RBM 算法的对比分析结果，由表中数据可知，RBM 只训练迭代了 20 次错误率就达到了 0.040293%，而 BP 算法训练迭代了 200 次，错误率还是高于只迭代了 20 次的 RBM，从而得知 RBM 的训练效果特别理想。

5. 结论

本文以云端海量大数据作为处理对象，通过融合 Hadoop 的大数据处理方案、AI 的 Deep Learning 相关技术以及数字水印的相关技术，实现了云端海量大数据的安全保护。采用 Map-Reduce 的并行计算框架机制，完成了海量大数据的计算，与此同时，提取云端海量大数据的标签，借用 HDFS 的分布式存储技术，实现了海量数据的分节点存储。利用 DBN 作为智能化的分类机制，对海量数据是否存在篡改进行检测判断，得到数据的分布式表示，以此实现了快速定位篡改区域以及对篡改数据的准确判断的目标。

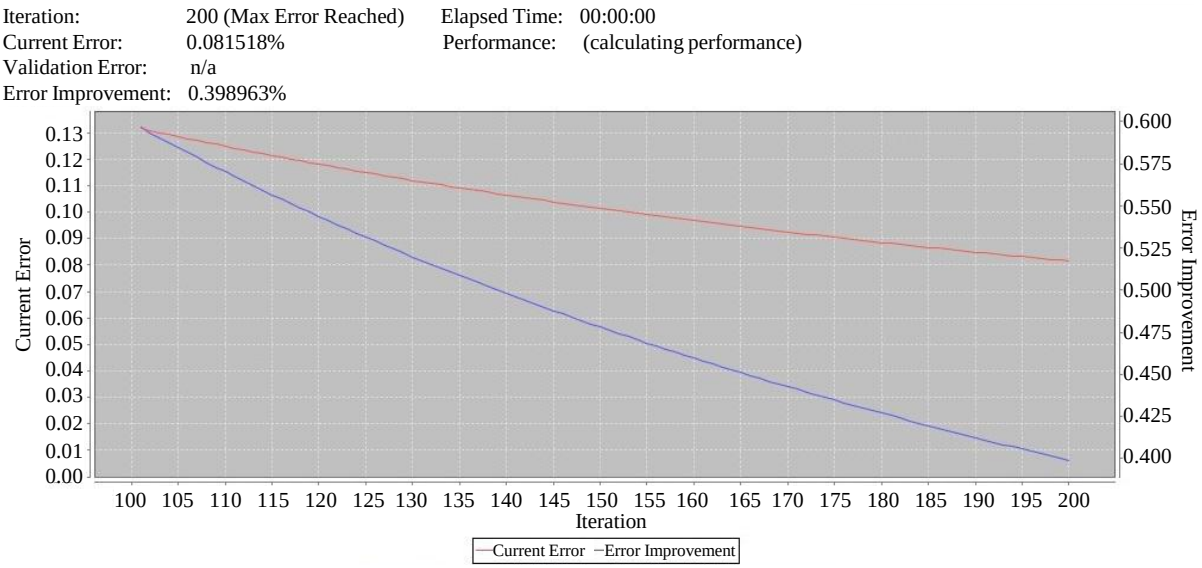

Figure 4. Performance curve of the BP algorithm
图 4. BP 算法训练性能曲线

Table 1. The training results of BP and RBM
表 1. BP 和 RBM 训练结果表

学习算法	训练迭代次数	当前的错误率	错误改善率
BP	200	0.081518%	0.398963%
RBM	20	0.040293%	0.440018%

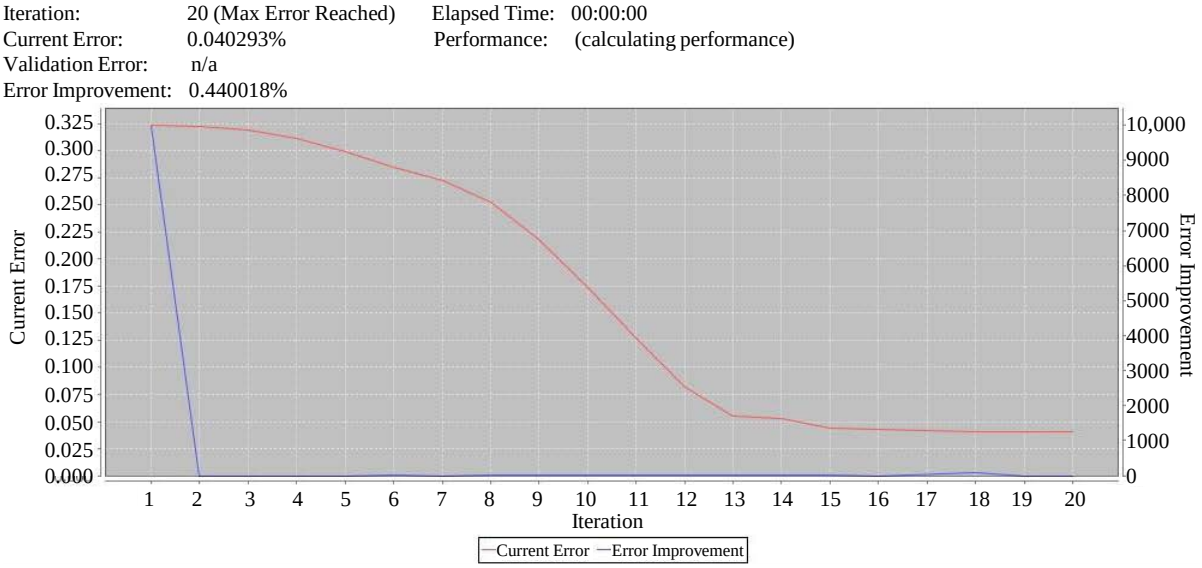

Figure 5. Performance curve of the RBM algorithm
图 5. RBM 算法训练性能曲线

基金项目

国家自然基金(61370145)；江苏省产学研联合创新项目(BY2013020)；徐州市科技计划项目(XM13B126)徐州工程学院青年基金(XKY2012309)。

参考文献 (References)

[1] Chen, T. and Liao, D.M. (2013) A study on fast post-processing massive data of casting numerical simulation on personal computers. *China Foundry*, **10**, 321-324.

[2] Zhou, P. and Dai, L.R. (2012) Combining information from multi-stream features using deep neural network in speech recognition. *Proceedings of* 2012 *IEEE* 11*th International Conference on Signal Processing* (*ICSP* 2012), Beijing, 21-25 October 2012, 557-561. http://dx.doi.org/10.1109/ICoSP.2012.6491549

[3] Hinton, G. (2013) Training Recurrent Neural Networks. Doctor of Philosophy Graduate, Department of Computer Science, University of Toronto, Toronto, 1-93.

[4] 朱珠 (2008) 基于 Hadoop 的海量数据处理模型研究和应用.北京邮电大学, 北京, 1-62.

[5] 刘鹏, 黄宜华, 陈卫卫 (2011) 实战 Hadoop——开启通向云计算的捷径. 电子工业出版社, 北京.

[6] 周长建, 司震宇, 等 (2013) 基于 Deep Learning 网络态势感知建模方法研究. *东北农业大学学报*, **5**, 144-149.

[7] 奚雪峰, 周国栋 (2014) 基于 Deep Learning 的代词指代消解. *北京大学学报*(*自然科学版*), **1**, 100-110.

[8] 谢桂兰 (2010) 基于 Hadoop Map Reduce 模型的应用研究. *微型机与应用*, **8**, 4-7.

[9] 田秀霞 (2011) 基于 Hadoop 架构的分布式计算和存储技术及其应用. *上海电力学院学报*, **1**, 70-74.