

基于 Spark 的支持隐私保护的聚类算法

高志强, 李庆鹏, 胡人远

(武警工程大学信息工程系, 陕西 西安 710086)

摘要: 针对经典聚类方法无法应对任意背景知识下恶意攻击者在海量数据挖掘过程中的恶意攻击问题, 结合差分隐私保护机制, 提出一种适用于 Spark 内存计算框架下满足差分隐私保护的聚类算法, 并从理论上证明了改进算法满足在 Spark 并行计算框架下的 ϵ -差分隐私。实验结果表明, 改进算法在保证聚类结果可用性前提下, 具有良好的隐私保护性和满意的运行效率, 在海量数据聚类分析的隐私保护挖掘中, 具有很好的应用前景和价值。

关键词: Spark; 差分隐私; 聚类算法; 数据挖掘; 大数据分析

中图分类号: TP301

文献标识码: A

doi: 10.11959/j.issn.2096-109x.2016.00087

Clustering algorithm preserving differential privacy in the framework of Spark

GAO Zhi-qiang, LI Qing-peng, HU Ren-yuan

(Department of Information Engineering, University of PAP, Xi'an 710086, China)

Abstract: Aimed at the problem that traditional methods fail to deal with malicious attacks with arbitrary background knowledge during the process of massive data clustering analysis, an improved clustering algorithm, especially designed for preserving differential privacy, under the framework of Spark was proposed. Furthermore, it's theoretically proved to meet the standard of ϵ -differential privacy in the framework of Spark platform. Finally, experimental results show that guaranteeing the availability of proposed clustering algorithm, the improved algorithm has an advantage over privacy protection and satisfaction in the aspect of time as well as efficiency. Most importantly, the proposed algorithm shows a good application prospect in the analysis of data clustering preserving privacy protection and data security.

Key words: Spark, differential privacy, clustering algorithm, data mining, big data analysis

1 引言

在大数据时代, 海量数据的分析和处理技术受到了越来越广泛的关注, 尤其是大型互联网公司, 数据就是商机, 数据就是价值。然而, 在开放系统环境下, 信息资源的共享在给用户带来极

大便利的同时, 数据的隐私与敏感信息的保护问题也不断凸显。用户的原始数据信息可能在数据挖掘和机器学习的某个过程中被恶意攻击者破坏、篡改、截获, 用户数据信息的隐私安全正面临着前所未有的威胁^[1]。因此, 支持任意背景知识下可以保护用户隐私信息的海量数据挖掘算法

收稿日期: 2016-05-17; 修回日期: 2016-06-26。通信作者: 高志强, 1090398464@qq.com

基金项目: 国家自然科学基金资助项目 (No.61309008); 陕西省自然科学基金资助项目 (No.2014JQ8049)

Foundation Items: The National Natural Science Foundation of China (No.61309008), The Natural Science Foundation of Shaanxi Province (No.2014JQ8049)

已成为研究热点^[2]。

国内外学者做了许多卓有成效的研究工作。文献[3]提出了一种基于 Haar 小波和 Daub-4 小波分析的数据变换方式,使经典的分类挖掘算法满足隐私保护的要求。文献[4]在朴素贝叶斯分类器的基础上,提出一种满足隐私保护的高效分布式推荐算法。文献[5]将隐私保护机制融入 Hadoop 平台下的 MapReduce 分布式计算框架,实现了海量分布式数据的隐私保护算法。

文献[6,7]均实现了 MapReduce 大数据并行计算框架与经典聚类算法的结合,其中,文献[6]主要针对聚类算法应对最大背景知识攻击的问题,保证了改进算法在挖掘过程中的隐私保护特性。文献[7]在 MapReduce 框架基础上,实现了支持差分隐私保护的 k -means 聚类算法,该算法在 Reduce 阶段为数值型变量 num 和 sum 添加 Laplace 噪声,实现了差分隐私保护。但是 MapReduce 框架在算法迭代过程中,需多次读写存储在外部的数据,消耗大量与 HDFS 的 I/O 通信资源,并且过多的噪声扰动也会增加隐私保护算法的复杂度开销。因此,为了提高对隐私数据的保护程度和挖掘结果的可用性,本文以海量数据的数理统计特性为出发点,提出一种 Spark 框架下能有效支持差分隐私保护的 k -means 聚类方法,使恶意攻击者即使在获取最大数据背景知识情况下,仍无法得到用户数据的敏感信息。

2 差分隐私保护

大数据时代,攻击者可以通过网络资源,利用多种黑客技术获取关于用户隐私数据的背景知识,从而导致了用户隐私的泄露,甚至引发犯罪和一系列更为严重的后果。而差分隐私保护技术(DP, differential privacy)^[8]通过向数据集加入与数据集大小无关的扰动噪声(DN, disturbance noise),降低原始用户数据在查询或结果分析中泄露的风险,以达到保护隐私的效果,其形式化定义如下。

定理 1 假设数据集 D_0 或 D_1 完全相同或只相差一条记录(D_0 和 D_1 被称为兄弟数据集),定义 F 为查询操作, $F(A)$ 为随机算法的查询输出, $\Pr[X]$ 为事件 X 发生的概率测度,如果对任意输出

$S \in F(A)$, 有

$$\Pr[A(D_0) \subseteq F(A)] \leq e^\epsilon \Pr[A(D_1) \subseteq F(A)] \quad (1)$$

则随机算法 A 为隐私预算 ϵ 提供 ϵ -差分隐私保护^[9]。

图 1 给出差分隐私算法的输出结果所对应的概率密度函数。

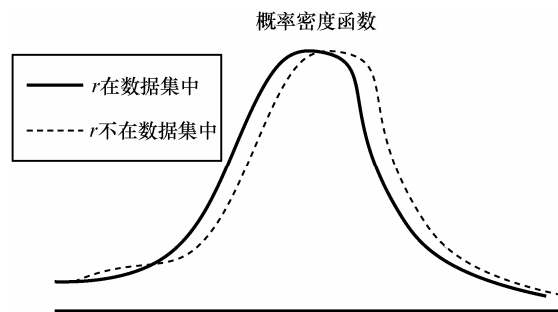


图 1 差分隐私算法输出的概率密度函数

由图 1 可知,单条记录 r 是否在兄弟数据集中,对满足差分隐私保护的随机算法的输出结果所对应的概率密度函数变化影响不大,也就是说,差分隐私保护算法所对应的输出函数值对于输入数据集的微小差异是不敏感的,攻击者无法轻易判断某条数据记录是否在数据集中。

对于查询函数 F ,全局敏感度是其重要固有属性,衡量单个记录的变化对查询函数输出的影响,其定义如下。

定义 1 查询函数 F 所对应的全局敏感度^[10]为

$$\Delta F = \max_{D_0, D_1} \|F(D_0) - F(D_1)\|_1 \quad (2)$$

其中,数据集 D_0 与 D_1 只相差一条记录, $\|\cdot\|_1$ 为向量的一范式。

本文采用了可以较好地保持原有数据统计特性的 Laplace 机制,使改进算法更适用于数值型聚类分析,其原理如图 2 所示。

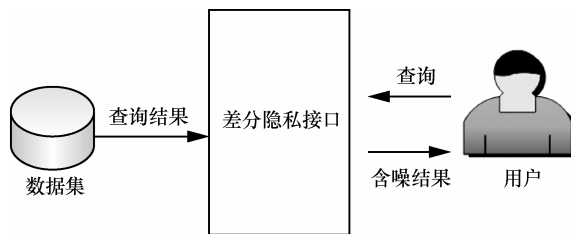


图 2 Laplace 机制

Laplace 机制通过向数值型数据中添加 Laplace 噪声实现差分隐私保护,其具体实现如定理 2。

定理 2 对于查询函数 F ，数据集 D_0 ，设查询输出为 $F(D_0)$ ， F 的全局敏感度为 ΔF ，如果噪声 Y 服从尺度为 $\frac{\Delta F}{\epsilon}$ 的 Laplace 分布，则算法 $A(D_0)=F(D_0)+Y$ 满足 ϵ -差分隐私^[11]。

随机扰动噪声 $\text{Lap}\left(\frac{\Delta F}{\epsilon}\right)$ 的概率密度函数为

$$p(x) = \frac{1}{2 \frac{\Delta F}{\epsilon}} \exp\left(-\frac{|x|}{\frac{\Delta F}{\epsilon}}\right) \quad (3)$$

此外，差分隐私保护具有序列组合性和并行组合性^[12]。

性质 1 设有 m 个随机算法 $A_1, \dots, A_m$ ，算法 $A_i (1 \leq i \leq m)$ 提供 ϵ_i -差分隐私保护，则对同一数据集 $D, \{A_1, \dots, A_m\}$ 在 D 上的序列组合算法提供 ϵ_i -差分隐私保护，其中

$$\epsilon = \sum_{i=1}^m \epsilon_i \quad (4)$$

性质 2 设有随机算法 A 和数据集 D ，将 D 分为不相交的子集 $D_1, \dots, D_n$ ，若算法 A 提供 ϵ -差分隐私保护，则 A 在 $\{D_1, \dots, D_n\}$ 上的组合算法所构成的算法提供 ϵ -差分隐私保护。

基于性质 1 和性质 2，本文证明了改进算法 DP HK-means 满足差分隐私并指导隐私预算分配过程。

3 Spark 框架下的 DP HK-means 算法

本文算法基于 Spark 内存计算框架，实现了海量数据在聚类分析时，在数据集的统计信息改变任一条记录时，具有最大背景知识的攻击者在整个聚类过程中仍无法获得数据隐私的功能，其对应的攻击模型如图 3 所示。

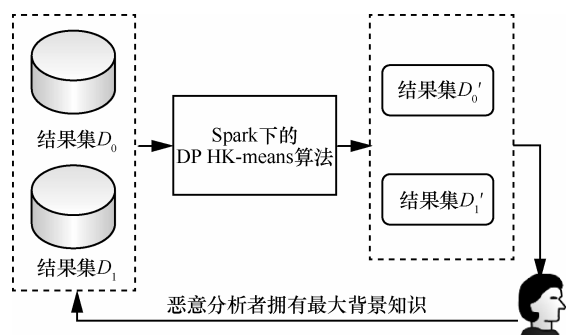


图 3 本文算法对应的攻击模型

3.1 算法设计

为解决 k -means 算法在聚类挖掘过程中不适合海量数据分布式挖掘、缺乏隐私保护功能等问题，本文提出一种适合于 Spark 内存迭代框架的 DP HK-means 算法，该算法主要执行步骤如下。

Step1 将从 HDFS 中读取的初始数据集转换为 Spark 框架下的 RDD 数据集，并对初始聚类中心点编码。

Step2 Map 操作。训练初始聚类中心，获得初始聚类中心，构成 (Key, Value) 键值对，并选择距离最小的聚类中心，判断记录的聚类归属。

Step3 Reduce 及隐私保护操作。计算各个聚类中记录数量 num 及其向量和 sum ，并对变量 sum 加入 Laplace 噪声扰动，重新计算相应聚类中心。

Step4 根据 Key 值，进行 Join 操作，产生新的 RDD 数据集，如果本轮和上轮聚类中心向量差小于预设值，则通过 SaveAsTextFile 方法保存聚类结果，结束迭代；否则循环 Step 2~Step 4，并将中间计算结果缓存至 Cache 中，进行下一轮操作。

Step5 输出聚类结果。

在 Step1 和 Step2 中，本文改进算法利用初始聚类中心改进初始中心的选择方式，加快聚类速率；在 Step3 中，由于变量 num 属于攻击者背景知识，为减少多余噪声对聚类中心的影响，只需对聚类内记录的向量和 sum 进行差分隐私处理以提高保护效率及聚类准确性；在 Step4 中，利用 Spark 框架的核心技术，将 RDD 缓存至 Cache，避免与 HDFS 的多次 I/O 通信，提高了算法执行效率。

3.2 隐私性分析

由于本文算法基于内存计算的每轮迭代过程均符合差分隐私的序列组合性质，所以，算法的 T 次迭代过程中隐私保护预算 ϵ 满足性质 1。另外，在本文采用的隐私预算分配方案中，每次迭代的隐私预算为 $\epsilon_t = \frac{\epsilon}{2^t}$ 。

由文献[7]的分析可知， sum 的全局敏感度为 $d+1$ (d 为数据集中所有记录的维数)。因此，在每轮迭代过程中添加大小为 $\text{Lap} \frac{(d+1)2^{t+1}}{\epsilon}$ 的 Laplace 噪声即可保证 Spark 框架下 DP HK-means 算法满足 ϵ -差分隐私。而且，本文算法只对 sum

中间变量加入噪声扰动,降低了隐私保护预算,提高了迭代效率,减少了过多噪声对聚类效果的影响。

4 实验与分析

实验平台搭建 5 个分布式节点,相关配置如下:操作系统为 CentOS 6.4, CPU 为 3.30 GHz, 内存为 16 GB, 320 GB 硬盘, 1 000 Mbit/s 以太网,用 Hadoop 2.6.4 和 Spark 2.0 搭建 Spark on Yarn 的集群环境^[13], 算法由 Python 实现。实验所选择的数据集为 UC Irvine Machine Learning Repository 中的 Adult Data Set 数据集(记录数为 48 842, 属性数为 14), 根据数据集的已知分类结果, 实验中聚类中心数优化为 6, 迭代阈值设置为 1。实验部分从运行效率及隐私保护处理后输出结果可用性方面与基于 MapReduce 的 DP *k*-means 算法^[7]进行对比测试。

4.1 算法效率分析

本文采用加速比(speedup)来测试算法的并行效率, 其中, T_s 为单机运行时间, T_c 为算法的集群执行耗时。

$$speedup = \frac{T_s}{T_c} \quad (5)$$

本文算法与对比算法的加速比结果如表 1 所示。

实验算法	加速比实验结果			
	加速比			
	2 节点	3 节点	4 节点	5 节点
本文算法	1.57	2.31	3.13	4.09
对比算法	1.14	1.48	1.62	1.83

由表 1 可知, 本文算法加速比明显优于对比算法^[7], 而且随着节点数量增大而增大, 虽然算法的加速比并没有按线性递增, 但也验证了 Spark 内存计算模式在迭代运算效率方面高于 MapReduce 框架。

4.2 算法聚类结果可用性分析

本文采用的聚类可行性指标为准确率与召回率的加权综合, 表达式为

$$F_{available} = \sum \frac{|U_i| F_i}{N} \quad (6)$$

其中, F_i 为准确率和召回率的加权调和平均值, $|U_i|$ 为第 i 个参考标准的聚类集合, N 为记录集记

录的总量。本文用 $F_{available}$ 来衡量聚类结果可用性, 衡量算法得到的聚类结果与参考结果的相似程度。当隐私预算改变时, 实验聚类结果可用性情况如图 4 所示。

由图 4 可知, 当隐私预算 ϵ 在 0~15 之间取值时, 聚类结果可用性属于 [0.3, 1.0]。当隐私预算水平较低时, 聚类结果可用性与其成很强的正相关, 当 ϵ 达到一定程度时, 可用性趋于平缓, 这也间接验证了 ϵ 对隐私性和可用性的平衡作用。

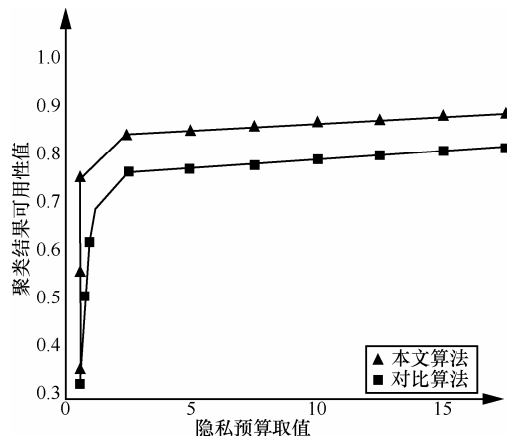


图 4 聚类可用性随 ϵ 变化情况

5 结束语

本文利用 Spark 并行内存计算技术实现了支持差分隐私的改进 *k*-means 算法, 改进算法优化算法优化初始聚类中心, 结合 Laplace 机制对聚类过程中最关键的隐私信息进行噪声扰动, 实现了 ϵ -差分隐私。同时, 在运行效率和聚类结果可用性方面, 改进算法性能明显优于传统基于 MapReduce 框架的聚类算法。今后研究工作将主要集中在保证隐私保护水平的前提下, 通过高效的保护机制来减少噪声的添加以降低算法复杂度。另外, 将群体智能优化算法融入 Spark 大数据计算框架, 也具有广阔的研究空间。

参考文献:

- [1] 方滨兴. 从层次角度看网络空间安全技术的覆盖领域[J]. 网络与信息安全学报, 2015, 1(1):1-6.
FANG B X. A hierarchy model on the research fields of cyberspace security technology[J]. Chinese Journal of Network and Information Security, 2015, 1(1):1-6.
- [2] 方滨兴, 贾焰, 李爱平, 等. 大数据隐私保护技术综述[J]. 大数据, 2016(1): 1-18.
FANG B X, JIA Y, LI A P, et al. A survey of large data privacy pro-

- tection technology [J]. Big Data, 2016, (1):1-18.
- [3] SANJAY B, ARYYA G. A Wavelet-based approach to preserve privacy for classification mining[J]. Decision Sciences, 2006, 37(4): 623-643.
- [4] CIHAN K, HUSEYIN P. Privacy-preserving naive Bayesian classifier-based recommendations on distributed data[J]. Computational Intelligence, 2015, 31 (1): 47-69.
- [5] ROY I, SETTY S T V, KILZER A, et al. Airavat: security and privacy for MapReduce[C]//The 7th Usenix Symposium on Networked Systems Design and Implementation. 2010:297-312.
- [6] 杨绍禹, 王世卿. MapReduce 模型下数据隐私保护机制研究[J]. 计算机科学, 2012, 39(12): 153-157.
YANG S Y, WANG S Q. Research on the mechanism of data privacy protection in the MapReduce model[J]. Computer Science, 2012, 39 (12): 153-157.
- [7] 李洪成, 吴晓平, 陈燕. MapReduce 框架下支持差分隐私保护的 k -means 聚类方法[J]. 通信学报, 2016, 37(2): 124-130.
LI H C, WU X P, CHEN Y. K -means clustering method for differential privacy protection under the framework of MapReduce[J]. Journal on Communications, 2016, 37 (2): 124-130.
- [8] DWORK C. A firm foundation for private data analysis[J]. Communications of the ACM, 2011, 54(1):86-95.
- [9] DWORK C. Differential privacy[C]//The 33rd International Colloquium on Automata, Languages and Programming. 2006: 1-12.
- [10] 熊平, 朱天清, 王晓峰. 差分隐私保护及其应用[J]. 计算机学报, 2014, 37(1): 101-122.
XIONG P, ZHU T Q, WANG X F. Differential privacy protection and its application [J]. Journal of Computer Science, 2014, 37(1): 101-122.
- [11] 丁丽萍, 卢国庆. 面向频繁模式挖掘的差分隐私保护研究综述[J]. 通信学报, 2014, 35(10): 200-209.
DING L P, LU G Q. A survey of differential privacy protection for frequent pattern mining[J]. Journal on Communications, 2014, 35 (10): 200-209.
- [12] 张啸剑, 孟小峰. 面向数据发布和分析的差分隐私保护[J]. 计算机学报, 2014, 37(4): 927-949.
MENG X F, ZHANG X J. Differential privacy preserving data publishing and analysis[J]. Journal of Computers, 2014, 37(4): 927-949.
- [13] 王保义, 王冬阳, 张少敏. 基于 Spark 和 IPPSO_LSSVM 的短期分布式电力负荷预测算法[J]. 电力自动化设备, 2016, 36(1): 117-122.
WANG B Y, WANG D Y, ZHANG S M. Short term distributed load forecasting algorithm based on Spark and IPPSO_LSSVM[J]. Electric Power Automation Equipment, 2016, 36(1): 117-122.

作者简介:



高志强 (1989-), 男, 河北故城人, 武警工程大学博士生, 主要研究方向为大数据隐私保护、群体智能优化算法。



李庆鹏 (1992-), 男, 山东济南人, 武警工程大学硕士生, 主要研究方向为大数据隐私保护、数据挖掘与机器学习。



胡人远 (1992-), 男, 浙江台州人, 武警工程大学硕士生, 主要研究方向为云计算数据存储、元数据管理、同态加密算法。