

# 面向垂直划分数据库的隐私保护分布式聚类算法

姚 瑶, 吉根林

(南京师范大学 数学与计算机科学学院, 江苏 南京 210097)

[摘要] 针对垂直划分的分布式数据库提出了一种基于隐私保护的分布式聚类算法 PPDC\_VP, 该算法基于 K-Means 的思想实现分布式聚类, 并且聚类过程中应用扰动技术保护本站点真实信息不被传送到其它站点, 从而达到隐私保护的目。理论分析和实验结果表明 PPDC\_VP 算法是有效的。

[关键词] 分布式聚类, 隐私保护, 扰动技术

[中图分类号] TP 311.1 [文献标识码] A [文章编号] 1672-1292(2008)04-0099-04

## Privacy-Preserving Distributed Clustering Algorithm Facing Vertically Partitioned Databases

Yao Yao, Ji Genlin

(School of Mathematics and Computer Science, Nanjing Normal University, Nanjing 210097, China)

**Abstract** Aiming at the vertically partitioned database, this paper presents a distributed clustering algorithm PPDC\_VP based on privacy-preserving. The algorithm is based on the idea of K-Means to realize distributed clustering and uses the perturbation technology to protect the real information of the site from being transferred to other sites in clustering procedure. Theoretical analysis and experimental results show that algorithm PPDC\_VP is effective.

**Key words** distributed clustering, privacy preserving, perturbation technology

近年来, 人们针对水平划分的数据库提出了一些分布式聚类算法, 如 K-Means<sup>[1]</sup>、DBDC<sup>[2]</sup>等, 这些算法不具有隐私保护功能, 它们在聚类过程中将本站点有关真实数据传送给其它站点, 从而导致信息泄露. 在实际分布式聚类应用中, 有时候需要保护本站点的真实信息不被传送给其它站点, 即需要进行隐私保护, 为此, 需要研究基于隐私保护的分布式聚类算法. 聚类过程中的隐私保护方法可大致分为数据扰动和安全多方计算两种. 基于数据扰动的隐私保护聚类思想是通过转换数据使得真实的敏感数据不为人知, 然后再进行聚类分析. 而基于安全多方计算的隐私保护聚类主要通过构造安全多方协议, 使得一组站点在仅仅拥有自己私有信息的情况下能最终获知全局聚类信息. 文献[3-4]提出应用平移、缩放、旋转等数据扰动方法进行集中式的隐私保护聚类. 基于安全多方技术的隐私保护挖掘具有代表性的算法是: 针对水平划分的分布式数据库提出了一种保持隐私的 EM 聚类的分布式算法<sup>[5]</sup>; 针对垂直划分的分布式数据库提出了一种基于安全交集的 K-Means 聚类<sup>[6]</sup>. 本文同样针对垂直划分的分布式数据库提出了一种基于隐私保护的分布式聚类算法 PPDC\_VP (Privacy Preserving Distributed Clustering Over Vertically Partitioned Database), 该算法基于 K-Means 的思想实现分布式聚类, 并且聚类过程中应用扰动技术保护本站点真实信息不被传送到其它站点, 从而达到隐私保护的目。理论分析和实验结果表明 PPDC\_VP 算法是有效的。

### 1 相关概念

#### 1.1 垂直划分数据库

**定义 1** 设分布式系统中有  $p$  个站点  $\{S_1, S_2, \dots, S_p\}$ , 各站点相应的局部数据集分别为  $\{DB_1, DB_2, \dots,$

收稿日期: 2008-06-18

基金项目: 国家自然科学基金(40771163)资助项目.

通讯联系人: 吉根林, 教授, 博士生导师, 研究方向: 数据挖掘技术及其应用. E-mail: glj@njnu.edu.cn

$DB_p\}$ , 全局数据集  $DB$  是  $\{DB_1, DB_2, \dots, DB_p\}$  的连接运算. 设  $DB$  中有  $n$  个对象, 分别为  $(x_1, x_2, \dots, x_n)$ , 这些对象用  $m$  维属性  $A = (a_1, a_2, \dots, a_m)$  描述, 站点  $S_i$  的数据集  $DB_i$  只拥有  $DB$  中的  $n_i$  个属性集合  $A_i, A_i \subset A$ . 这种划分称为垂直划分, 又称异质划分.

定义 2 设全局数据集  $DB$  可划分为  $k$  个聚簇  $\omega_1, \omega_2, \dots, \omega_k$ , 每个簇中的数据点个数分别为  $t_1, t_2, \dots, t_k$ , 聚簇  $\omega_j (j = 1, 2, \dots, k)$  所对应的各站点的局部聚类中心为  $\{c_{1j}, c_{2j}, \dots, c_{pj}\}$ , 站点  $S_i$  中, 对象  $x_g$  到局部中心点  $c_{ij}$  的距离称为局部距离  $d_{ij}$ , 其中  $d_{ij} = |x_g - c_{ij}|$ . 对象  $x_g$  在聚簇  $\omega_j (j = 1, 2, \dots, k)$  中所对应的全局距离  $D_j = (d_{1j} + d_{2j} + \dots + d_{p-1j} + d_{pj})$ , 聚类的目标函数  $E = \sum_{i=1}^k \sum_{j=1}^{t_i} d_{ij}(x_j, c_i)$ , 其中  $d_{ij}(x_j, c_i)$  是数据点  $x_j$  和中心点  $c_i$  之间的距离.

1.2 扰乱技术

扰乱技术的关键是解决具体扰乱方法与现有挖掘算法的有效整合. 扰乱技术的一种方法是数据交换, 通过交换不同记录之间的数值来隐藏记录所属对象与数值间的对应关系, 即改变数据的先后次序但不改变它们的数值. 扰乱技术的另一种方法是随机响应, 为数据增加噪声以保护真实数据不被发现. 通过对数据的随机化处理, 伪装数据, 使数据无法被对方知晓.

2 算法设计与描述

2.1 最近簇计算

K-Means 算法的重点是如何判断最近簇, 所谓最近簇是全局对象与哪一类中心的距离最近, 此类就是最近簇. 在分布式聚类中, 最近簇需要依靠全局距离进行判断. 若未加隐私保护, 主站点  $S_p$  计算对象  $x_g$  在聚簇  $\omega_j (j = 1, 2, \dots, k)$  中所对应的全局距离计算公式为:  $D_j = (d_{1j} + d_{2j} + \dots + d_{(p-1)j} + d_{pj})$ , 其中的最小值  $D_{min\_cluster}$  所对应簇称为最近簇, 并将  $x_g$  划分给最近簇.

加隐私保护后, 每个从站点  $S_i$  产生一个随机值  $R_i$  扰乱各局部距离  $d_{ij}$  为  $d'_{ij} = d_{ij} + R_i$ , 传送经过扰乱后的局部距离  $d'_{ij}$  给主站点  $S_p$  计算全局距离. 全局距离计算公式为:

$$D'_j = d'_{1j} + d'_{2j} + \dots + d'_{(p-1)j} + d_{pj} = (d_{1j} + R_1) + (d_{2j} + R_2) + \dots + (d_{(p-1)j} + R_{p-1}) + d_{pj} =$$
$$(d_{1j} + d_{2j} + \dots + d_{(p-1)j} + d_{pj}) + (R_1 + R_2 + \dots + R_{p-1}) = D_j + \sum_{i=1}^{p-1} R_i$$

所有全局距离  $(D_1, D_2, \dots, D_k)$  经过同一扰乱值  $\sum_{i=1}^{p-1} R_i$  扰乱后变为  $(D'_1, D'_2, \dots, D'_k)$ , 其大小关系并未改变, 按照  $(D'_1, D'_2, \dots, D'_k)$  所求得最近簇与按照  $(D_1, D_2, \dots, D_k)$  所求得最近簇相同. 因此传到主站点经过扰乱的数据无需经过解密就可以判断最近簇, 保证了最近簇判断的正确性.

2.2 算法思想

在通信过程中, 需要保护全局对象到局部类中心的距离. 利用主站点收集各从站点的局部距离, 从而计算全局对象到聚类中心的全局距离. 同时判断对象属于哪一类. 因为各从站点发送给主站点的局部距离经过统一扰乱, 所以可保证在不泄露各自私有数据的情况下, 得到与 K-Means 等效的聚类结果.

不失一般性, 令  $S_p$  为主站点,  $S_1, \dots, S_{p-1}$  为从站点. 首先,  $S_p$  任意选择  $k$  个对象作为本方初始的类中心, 并把对象编号发送给  $S_1, \dots, S_{p-1}$ . 从站点  $S_i$  接收到编号后, 初始化本站点聚类中心  $\{c_{1i}, c_{2i}, \dots, c_{ki}\}$ , 并且随机生成扰乱值  $R_i$ . 计算站点  $S_i$  中每个对象  $x_g$  到各聚类中心距离  $d_{ij}$ . 利用扰乱值  $R_i$  伪装各中心距离  $d_{ij}$  为  $d'_{ij}$ . 发送经过扰乱后的局部聚类中心距离  $(d'_{1i}, d'_{2i}, \dots, d'_{ki})$  给主站点  $S_p$ .  $S_p$  接收到各从站点局部距离后, 计算  $x_g$  到各类中心的全局距离  $(D'_1, D'_2, \dots, D'_k)$ , 找出最小的距离  $D'_{min\_cluster}$ , 其中下标  $min\_cluster$  即是对象  $x_g$  在此次迭代中所属的类. 广播  $min\_cluster$  给各从站点, 划分  $x_g$  到所属聚簇. 当所有对象执行完一次聚类后, 重新计算聚类中心, 进行迭代, 直到  $E$  稳定, 算法结束.

2.3 PPDC-VP 算法描述

输入: 局部数据集  $\{DB_1, DB_2, \dots, DB_p\}$ , 聚簇的个数  $k$

输出:  $k$  个聚簇

步骤:

```
Master site  $S_p$ :
    initialize (  $c_{11}, c_{12}, \dots, c_{1k}$  );
    broadcast ( index(  $k$  ) );
Slave site  $S_i(1 \leq i < p)$ :
    receive ( index(  $k$  ) );
    initialize (  $c_{i1}, c_{i2}, \dots, c_{ik}$  );
while  $E$  is not stable do
{Master site  $S_p$ :
    for each data object  $x_g \in DB_p (g = 1, 2, \dots, n)$  do
    { for  $i = 1$  to  $p - 1$  do
        receive (  $d'_{i1}, d'_{i2}, \dots, d'_{ik}$  );
        for  $j = 1$  to  $k$  do
            {  $d_{pj} = \text{dist\_Comput}(x_g, c_{pj})$ ;
               $D'_j = (d'_{1j} + d'_{2j} \dots + d'_{(p-1)j} + d_{pj})$ ; }
             $D'_{\text{min\_cluster}} = \text{geM\_in}(D_1, D_2, \dots, D_k)$ ;
            partition (  $x_g, \text{min\_cluster}$  );
            broadcast ( min\_cluster ); }
        computer (  $c_{p1}, c_{p2}, \dots, c_{pk}$  );
        computing (  $E$  );
Slave site  $S_i(1 \leq i < p)$ :
    for each data object  $x_g \in DB_i (g = 1, 2, \dots, n)$  do
    {  $R_i = \text{random}()$ ;
      for  $j = 1$  to  $k$  do
          {  $d_{ij} = \text{dist\_Comput}(x_g, c_{ij})$ ;
             $d'_{ij} = d_{ij} + R_i$ ; }
          send (  $d'_{i1}, d'_{i2}, \dots, d'_{ik}$  ) to Master site;
          receive ( min\_cluster );
          partition (  $x_g, \text{min\_cluster}$  );
          computer (  $c_{i1}, c_{i2}, \dots, c_{ik}$  );
    }
```

3 实验结果与分析

我们使用 3 台微机构成 100Mb 的局域网, 微机配置为 IntelPentium IV 2.93GHz/512MB, 开发环境为 JBuilder 2006 Enterprise, 利用 Java 实现了 PPDC\_VP 算法. 实验数据来自 UCI 机器学习数据库<sup>[7,8]</sup>中, 实验数据源如表 1 所示. 各局部站点分别从统一数据集中抽取相应属性值. PPDC\_VP 算法的聚类精度与传统的集中式 K-Means 算法作比较, 结果如图 1 所示. 实验表明 PPDC\_VP 的聚类效果与传统集中式 K-Means 聚类效果相当. PPDC\_VP 算法的效率与未加隐私保护的面向垂直划分数据库的分布式聚类算法 (本文命名为 DC\_VP) 进行比较, 在数据集 glass 与 segmentation test 的运行效率如图 2 所示, 实验表明在同等数据量下, 在 segmentation test 比 glass 数据集上的运行时间稍长, 其原因是 segmentation test 分布在各站点的属性较多, 而且类别个数也多, 其计算和通信量大.

表 1 实验数据集  
Table 1 Experiment data sets

数据集	对象个数	属性维数	类别个数	属性分割情况		
				1 站点	2 站点	3 站点
glass	200	9	6	3	3	3
segmentation test	200	17	7	6	6	5
segmentation	2 000	17	7	6	6	5

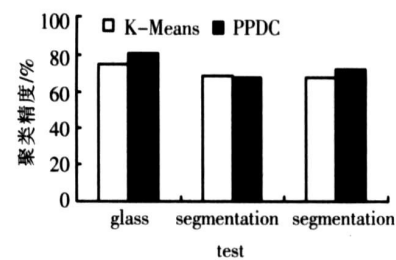


图 1 PPDC\_VP 与 K-Means 的聚类精度比较  
Fig.1 The accuracy of cluster comparison

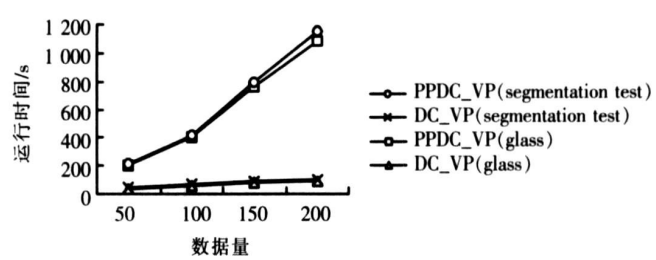


图 2 PPDC\_VP 与 DC\_VP 算法效率比较  
Fig.2 The algorithm efficiency comparison

4 结语

本文针对垂直划分的分布式数据库提出的分布式隐私保护聚类算法 PPDC\_VP采用扰乱技术,伪装各从站点所要通信的局部距离信息,不仅能够达到隐私保护的目的,而且具有较高效率.扰乱技术同样适合于水平划分的分布式数据库隐私保护聚类,研究与实现水平划分的分布式隐私保护聚类算法将是我们下一步的工作.

[参考文献] (References)

[ 1 ] Kantabutra S, Couch A L. Parallel k-means clustering algorithm on Now[s][J]. NECTEC Technical Journal, 2000, 1(6): 243-247.

[ 2 ] Januzaj E, Kriege I H P, Pfeifle M. DBDC: density based distributed clustering[C] // Proc the 9th Int'l Conf Extending Database Technology, Heraklion, Greece, Springer, 2004, 88-105.

[ 3 ] Stanley R M, Oliveira Osmar R, Zaiane. Privacy preserving clustering by data transformation[C] // Proc of the 18th Brazilian Symposium on Databases, Manaus, Brazil, Springer, 2003, 304-318.

[ 4 ] Stanley R M, Oliveira Osmar R, Zaiane. Achieving privacy preservation when sharing data for clustering[C] // Proc of the International Workshop on Secure Data Management in a Connected World, Toronto, Canada, Springer, 2004, 67-82.

[ 5 ] Lin X, Clifton C, Zhu M. Privacy preserving clustering with distributed EM mixture modeling[J]. Knowledge and Information Systems, 2005, 8(1): 68-81.

[ 6 ] Vaidya J, Clifton C. Privacy-preserving K-means clustering[C] // The 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington DC, USA: ACM Press, 2003, 593-599.

[ 7 ] B German. Glass Identification Database[DB/OL]. 1987[2007] <http://machine.uci.edu/databases>

[ 8 ] Vision Group. Image Segmentation data[DB/OL]. 1990[2007] <http://machine.uci.edu/databases/statlog/segment>

[责任编辑: 刘健]