

基于分布式混合数据挖掘的电信客户流失分析

李爱群^{1,2}, 乔 晗¹, 王汝传^{1,2}, 邓 松¹

(1. 南京邮电大学 计算机学院, 江苏 南京 210003;

2. 南京邮电大学 计算机研究所, 江苏 南京 210003)

摘 要: CORBA 技术庞大而复杂, 且技术和标准的更新相对较慢。电信运营企业应用系统是客户流失分析的主要数据来源, 而传统的客户流失分析由于该系统数据的集中式存储继而采用集中式挖掘, 对海量数据的挖掘效率低下。为进一步提高挖掘效率, 提出网格下基于分布式混合数据挖掘的电信客户流失分析 (Customer Churn Analysis upon Distributed Hybrid Data Mining in Grid, CCA-DHDM), 并借助 GridSphere 门户, 在该平台上实现了 BP 神经网络算法和 K-Means 聚类算法。仿真实验表明, 与单机环境相比, 随着网格节点数增加, 算法的平均耗时明显下降 65% 到 75%, 同时算法的效率得以较明显地提高。

关键词: 客户流失分析; 网格计算; BP 神经网络; K-Means 聚类算法

中图分类号: TP39

文献标识码: A

文章编号: 1673-629X(2010)10-0043-04

Telecommunication Carriers Customer Churn Analysis Based on Distributed Hybrid Data Mining

LI Ai-qun^{1,2}, QIAO Han¹, WANG Ru-chuan^{1,2}, DENG Song¹

(1. College of Computer, Nanjing University of Posts and Telecommunications, Nanjing 210003, China;

2. Institute of Computer, Nanjing University of Posts and Telecommunications, Nanjing 210003, China)

Abstract: CORBA is a large and complex technology, and the updating of technique and standard is relatively slow. Telecom enterprise application systems are the main source of data for customer churn analysis, the traditional customers churn analysis uses a centralized mining due to centralized data storage, and the mining efficiency for mass data is low. Present CCA-DHDM (Customer Churn Analysis upon Distributed Hybrid Data Mining in Grid), and achieve the BP neural network algorithm and K-Means clustering algorithm in this platform by means of GridSphere Portal. The simulation shows that, compared with stand-alone environment, the average time-consuming of algorithm is decreased obviously 65 percent to 75 percent with the grid nodes increasing, and the efficiency of the algorithm is improved clearly.

Key words: customer churn analysis; grid computing; BP neural network; K-Means clustering algorithm

0 引 言

当前, 国内电信运营企业应用系统具有多操作系

统、多数据库、数据量大、网络环境异构以及业务应用不同等特点。而这些系统又是电信客户流失分析^[1,2]的数据来源, 因此具有天然的分布性。传统的方法采用集中式进行处理, 把所有地理分散的数据集中传输到指定的服务器上, 然后对其进行客户流失分析, 这样由于传输的数据量过于庞大, 一方面造成一定的存储压力, 另一方面会造成很大的通信开销, 而且对海量数据进行集中式挖掘, 挖掘的效率低下, 因此需要新技术来实现高性能的分布式电信客户流失分析。

传统的分布式电信客户流失分析较多地采用 CORBA 中间件^[3]技术。CORBA 技术具有良好的互操作性和开放性。但 CORBA 技术庞大而复杂, 且技术和标准的更新相对较慢, 而网格计算^[4,5]由于其强大的分布式计算能力^[6,7]、较强的扩展性^[8]以及易实现

收稿日期: 2010-02-08; 修回日期: 2010-06-20

基金项目: 国家自然科学基金 (60973139, 60903181, 60773041); 江苏省自然科学基金 (BK2008451); 省级现代服务业发展专项资金 (2010002); 江苏省高校自然科学基金基础研究项目 (09KJB520009); 国家和江苏省博士后基金 (0801019C, 20090451240, 20090451241); 江苏高校科技创新计划项目 (CX09B-153Z, CX08B-086Z); 江苏省六大高峰人才项目 (2008118); 江苏省计算机信息处理技术重点实验室基金 (2010)

作者简介: 李爱群 (1969-), 女, 浙江海宁人, 讲师, 研究方向为计算机系统、计算机网路、智能计算等; 王汝传, 教授, 博士生导师, 研究方向为计算机软件、计算机网路和网路、对等计算、信息安全、无线传感器网路、移动代理和虚拟现实技术等。

的优势^[9]已成为现在分布式计算平台的首选^[10]。

随着网格这种分布式技术的不断发展,利用网格平台可以很好地处理电信客户流失分析^[11,12],为此文中提出了网格下基于分布式混合数据挖掘的客户流失分析系统(Customer Churn Analysis upon Distributed Hybrid Data Mining in Grid, CCA-DHDM),试验证明,通过构建网格平台来对地理位置分散的各种数据进行分布式挖掘处理,一方面大大降低了存储压力,另一方面提高了挖掘的效率和准确率。

文中所做的贡献主要为:

(1)为了更好地对电信客户数据进行分类,提出基于 BP 神经网络的客户数据分类算法(Classification of Customer Data on BP);

(2)提出基于 K-Means 的客户流失原因聚类算法(Clustering of Customer Churn Reason on K-Means),通过该聚类算法把流失客户的原因进行分析,从而为决策者提供相应的策略;

(3)进行比较实验,依实验数据予以性能分析。

1 基于网格服务的电信客户流失分析

1.1 系统框架

本系统的大体框架分为服务端和客户端。服务端有整个客户流失分析各功能所需要的服务:BP 神经网络分类服务和 K 均值聚类分析服务,客户端在需要调用服务时,向服务端发出请求。经过对数据的预处理后,基于网格平台之上,分多个节点对数据予以 BP 神经网络分类,并得出结果。之后,针对 BP 神经网络分类结果中流失数据再予以 K 均值聚类分析。完成 K 均值聚类分析,对最终结果予以汇聚。最后将处理出来的结果反馈给客户端。具体结构图如图 1 所示。

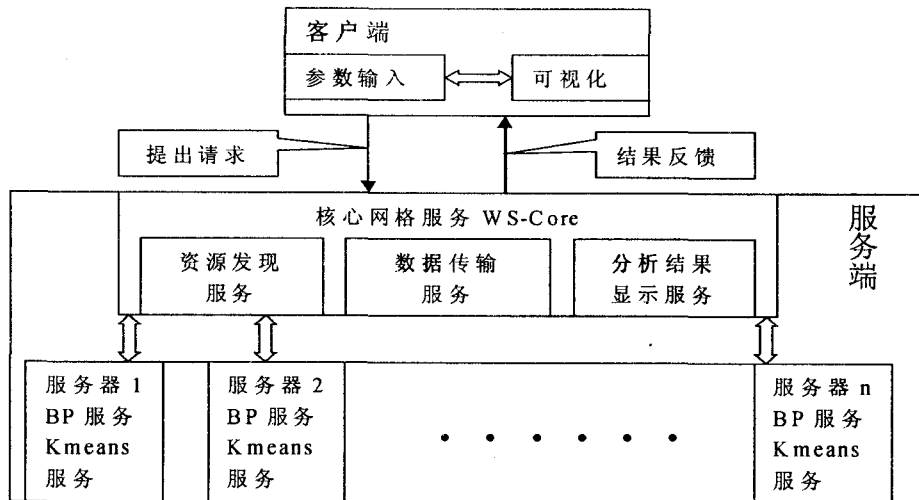


图 1 基于分布式混合数据挖掘的
电信客户流失系统框架图

1.2 具体实现

算法 1:BP 神经网络算法

●服务端

1. Receive (InputCd, OutputCd, StEfficient, Bpecho, BpPrecision, BPGSAd, i, Sample[i]); //从客户端接收 BP 算法的参数、具体的 BP 算法服务地址以及待分类的样本数据块名等各种参数;

2. Initial (); //初始化 BP 网络的结构;

3. int AllNum = Statistics (Sample); //统计待分类样本数据的个数;

3. while (echo < Bpecho) {

4. 计算隐层、输出层的纯输入和输出;

5. 计算输入层到隐层以及隐层到输出层的误差;

6. BP 网络结构的更新;

7. ClassifiedNum ++; } //统计被正确分类的样本个数;

8. double BpPrecision [i] = ClassifiedNum / AllNum; //统计局部分类精度。

●客户端

Input: BP 算法服务地址 BPGSAd; 待分类的样本数据 Sample[i]; 输入层节点数(数据文件属性个数) InputCd, 输出层节点数(分类个数) OutputCd, 学习效率 StEfficient; 算法迭代次数 Bpecho; 算法精度 BpPrecision;

Output: 分类精度 ClassfiPrecision;

隐层节点数 HideCd 直接由输入层节点数 $\times 2 +$ 输出层节点数得到。

int gridcodes = SelectGridCodes (); //选择部署 BP 算法服务的网格节点;

1. for (int i=0; i < gridcodes; i++) {

2. String Sample[i] = Partition(gridcodes, data); //对待分类的样本数据根据选择网格节点数进行分割;

3. DataTransService(i, Sample[i]); //通过数据传输服务把分割好的数据块传送到指定的网格节点处;

4. Receive (InputCd, HideCd, OutputCd, StEfficient, Bpecho, BpPrecision, BPGSAd, i, Sample[i]); //传递服务端所需的参数;

5. double precision += DataTransService (BpPreci-

sion[i]); //把第 i 个网格节点所得到的局部分类精度返回到客户端;

6. ClassfiPrecision = precision/gridcodes; //计算全局分类精度;

7. Show (ClassfiPrecision); //返回分类精度并可视化显示。

BP 算法客户端在得到上述客户输入的数据,判断数据是否为空后,在不为空的情况下分别将上述数据传至服务地址中所选的服务器上。在服务器运算完 BP 算法之后得到分类精度。从上述算法描述中可见,整个基于网络的 BP 神经网络算法的执行时间主要包括数据的传输时间以及 BP 神经网络算法对于局部数据块的分类时间,并且网格下执行 BP 算法的时间复杂度并没有改变。这一点在下一小节的仿真实验中表现明显,局域网环境下,随着待处理数据的数据量和复杂度的增大,相比较而言,数据分块后传输到各网格节点的时间比 BP 算法处理整个海量数据时间要小的多。

算法 2: K-Means 算法

● 服务端

1. Receive (File3, K, InputParamt, KGSAAd, Sample[i]); //从客户端接收 K-Means 算法的参数、具体的 K-Means 算法服务地址以及待聚类的样本数据块名等各种参数;

2. Initial (); //初始化 K-Means 算法参数;

3. setInitialClusterCenter(); //随机产生初始聚类中心位置

4. while (i < SampleNum) {

5. cal_Distance(i, a[j]); 计算每个对象与每个聚类对象的均值(中心对象)的距离;

6. MinValue = sum[i][a[j]]; 根据最小距离重新对相应对象进行划分;

7. clusterNum ++; //统计聚类次数;

8. cluster[m].addElement(new Integer(i)); //记录聚类结果。

● 客户端

Input: K-Means 算法服务地址 KGSAAd; 待聚类数据 Sample[i]; 聚类个数 K, 待聚类数据属性个数 InputParamt;

Output: 聚类个数 CluNum; 聚类结果 CluRst;

int gridcodes = SelectGridCodes (); //选择部署 BP 算法服务的网格节点;

1. for (int i = 0; i < gridcodes; i++) {

2. String Sample[i] = Partition(gridcodes,

data); //对待聚类的样本数据根据选择网格节点数进行分割;

3. DataTransService (i, Sample[i]); //通过数据传输服务把分割好的数据块传送到指定的网格节点处;

4. Receive (File3, K, InputParamt, KGSAAd, Sample[i]); //传递服务端所需的参数;

5. double cluster + = DataTransService (cluster[m]); //把第 i 个网格节点所得到的聚类结果返回到客户端;

6. Show (cluster[m]); //返回聚类结果并可视化显示。

K-Means 聚类算法客户端在得到上述客户输入的数据,判断数据是否为空后,在不为空的情况下分别将上述数据传至服务地址中所选的服务器上。在服务器运算完 K-Means 聚类算法之后得到聚类结果。从上述算法描述中可见,整个基于网络的 K 均值聚类算法的执行时间主要包括数据的传输时间以及 K 均值聚类算法对于局部数据块的聚类时间,并且网格下执行 K 均值聚类算法的时间复杂度并没有改变。这一点在下一小节的仿真实验中表现明显,局域网环境下,随着待处理数据的数据量和复杂度的增大,相比较而言,数据分块后传输到各网格节点的时间比 K 均值聚类算法处理整个海量数据时间要小的多。

2 仿真实验和分析

为了验证在基于网格服务开发客户流失分析的可行性和有效性,文中在实验室局域网的环境下做了电信客户流失分析的实验。整个实验平台为 Windows XP + WS-Core-4.0.2 + Jdk1.5 + Eclipse3.1 + Tomcat5.0.28 + Gridsphere2.2.9 + Ant1.6.5,所有的程序由 Java 语言实现。

实验:将四个服务器和一台 pc 机连接在一个局域网中,pc 机作为客户端,四台服务器作为服务端。服务器上配置文中所需的所有所需的 BP 神经网络服务和 K-Means 聚类服务。客户端仅仅安装网络浏览器。客户端界面输入需要相应的参数,并提交分析请求。图 2 是单机情况运行和多节点情况下运行两个服务的

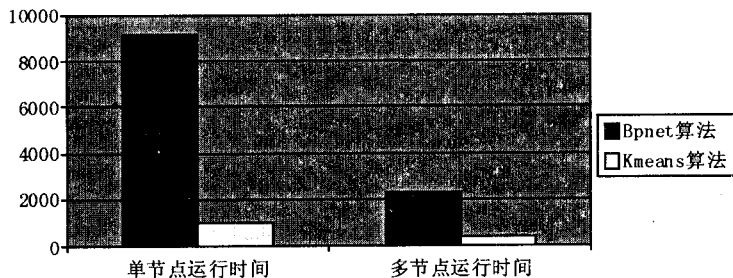


图 2 单节点与多节点运行服务耗时对比

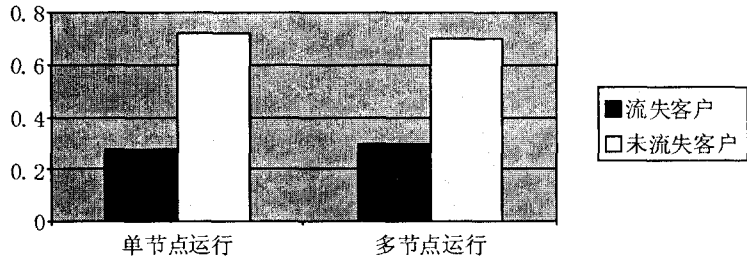


图 3 单节点与多节点运行 BP 神经网络服务结果对比图

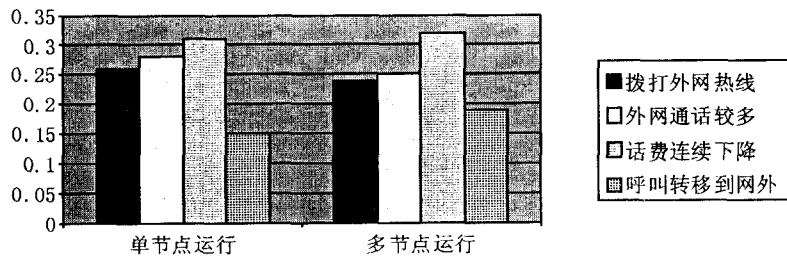


图 4 单节点与多节点运行 K-Means 聚类服务结果对比

耗时对比。图 3 和图 4 分别是单节点和多节点运行 BP 神经网络服务和 K-Means 聚类服务的结果对比。

由于本系统是基于网格平台的电信客户流失分析，因此无论是从运行速度上还是运行的准确性上都有不同程度的提高。该实验结果以客户端图形化界面为呈现，分别在单服务节点和多服务节点的情况下，运行电信客户流失分析所需要的两个服务程序。并将运行耗时和运行结果予以比较。由图 2 看出，多节点运行 BP 算法服务时间比单节点运行下降 75%，多节点运行 K-Means 算法服务时间比单节点运行下降 65%。

由定义 1 并结合图 3，在 BP 算法服务运行结果中，单节点运行的准确率为 87.13%，多节点运行的准确率为 90.37%，相比而言准确率略有提升。由定义并结合图 4，在 K-Means 算法运行结果中，单节点运行的 K-Means 算法准确率为 60.37%，多节点运行行为 67.83%，相比而言聚类效果稍好。详见表 1。

表 1 算法各参数比较

	单节点	多节点
BP 算法运行时间	9103	2150
K-Means 算法运行时间	975	368
BPNN CV	87.13%	90.37%
K-Means 算法准确率	60.37%	67.83%

由表 1 可知，CCA-DHDM 算法在保证算法运行准确率甚至略有提升的基础上，提高了客户流失分析的效率，从而达到了电信客户流失分析的目的。

3 结束语

本系统基于网格实现了电信客户流失分析的成功尝试。在服务端，有用于客户流失分析的两个服务：BP 神经网络分类服务和 K-Means 聚类分析服务，并发布于网格上。在客户端，如需对相关客户信息进行处理时，借助 GridSphere 门户只需以 Web 网页的形式予以访问，然后依次调用网格上的上述两个服务，运算之后得出结果。本系统的设计充分考虑到了用户数据的分布性，因此全面地发挥了网格的优势，加快了数据处理的速度，提高了运行的准确性。

参考文献：

- [1] 李志刚. 客户关系管理理论与应用[M]. 北京:机械工业出版社, 2006.
- [2] Chris R, Jyun-Cheng W, David C Y. Data mining techniques for customer relationship management[J]. Technology in Society, 2002, 24(4):483-502.
- [3] 李琪林,刘 强,周明天. 论中间件技术及其分类[J]. 四川师范大学学报(自然科学版), 2001(6): 31-32.
- [4] Joseph, Joshy. Grid computingp[M]. Upper Saddle River, N. J.:Prentice Hall Professional Technical Reference, 2004.
- [5] Talia D, Trunfio P, Verta O. WSRF Services for Composing Distributed Data Mining Applications on Grids: Functionality and Performance[J]. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2006(2):1080-1089.
- [6] 赵冬梅,刘海峰,刘晨光. 基于 BP 神经网络的信息安全风险评估[J]. 计算机工程与应用, 2007(1): 41-42.
- [7] 赵小会. 函数逼近的发展——神经网络[J]. 中国科技信息, 2008(11): 22-23.
- [8] 洪仁植,王树达,常 亮. 基于 BP 神经网络的管道缺陷模式识别与精确定量识别[J]. 大庆石油学院学报, 2008(1): 23-24.
- [9] 郑玉明,史晶蕊,廖湖声. 文本分类的神经网络模型[J]. 计算机工程, 2005(21): 15-16.
- [10] 王向阳,于雁春. 基于改进 K-均值聚类的快速分形图像编码算法[J]. 计算机科学, 2008(2): 63-64.
- [11] 王 强,皮德常,李伟奇,等. 基于 Agent 和数据挖掘的分布式信息审计平台[J]. 计算机技术与发展, 2006,16(4): 141-143.
- [12] 侯敬军,曾致远,向 凌. 一种基于 Web 服务的分布式数据挖掘体系结构[J]. 微机发展(现更名:计算机技术与发展), 2004,14(6):48-51.