Data Mining Based on Clouds Computing

HAN Bo[1], ZHANG Rong-li[2]
***Shangluo University, Shaanxi Shangluo,China,* 726000**
1. *hanbo*3000@163.*com*, 2. *jsjxzhangrongli*@163.*com*

Abstract: Data Mining Based on Clouds Computing is the main business models in recent years, research under the cloud-based data mining technology is a new challenging issues, The discusses cloud computing cluster based on the data mining service model method, introduced in the system's equipment, raw data, fill data, fuzzy and data reduction, rule extraction, algorithm description, etc., before and after comparison of the algorithms of data changes in the rule base reflects the system model based on cloud computing services descriptive data under the general rules of adaptation and the expression of accurate and refined.

Keywords: Clouds Computing; Rule; Data Mining；Knowledge base

基于云计算集群服务的数据挖掘

韩 波[1]， 章荣丽[2]
1.商洛学院，陕西商洛，中国，726000
1. hanbo3000@163.com, 2. jsjxzhangrongli@163.com

【摘要】基于云计算的集群服务是近几年来主要的商业运行模式，研究基于云计算下数据挖掘的技术是一个全新的具有挑战性的课题，本文探讨了了基于云计算的集群服务模式的数据挖掘方法，介绍了系统中原始数据的装备、填入数据的模糊及数据的消减、规则的提取、算法描述等等，比较了算法前后规则库中数据的变化，体现了系统对基于云计算服务模式下数据描述性的普遍适应性和规则研究表达的准确和精炼性。

【关键词】云计算；规则；数据挖掘；知识库

1. 云计算概念

云计算是近几年提出的一种全新的商业计算模型。它将计算任务分布到由大量计算机构成的资源池上，从而使用户能够根据需要获取计算能力、存储空间和信息服务,这种资源池称为“云”[1]。

云计算（Cloud Computing）是网格计算（Grid Computing）、分布式计算（Distributed Computing）、并行计算（Parallel Computing）的延伸，或者说是这些计算科学概念的完美融合[2]。之所以称其为云，主要是因为它在某些方面具有现实生活中云的特征：云一般都较大，其规模可以动态延伸而且边界模糊不定，云在空中飘忽不定，你无法也无需确定它的具体位置，但它确实存在于某处。因此“云”是一些可以自我维护和管理的虚拟计算资源，通常为一些大型服务器集群，包括计算服务器、存储服务器、宽带资源等等。云计算将所有的计算资源集中起来，并由软件实现自动管理，无需人为参与。这使得应用提供者无需为繁琐的细节而烦恼，能够更加专注于自己的业务，有利于创新和降低成本。

那么对基于云计算下的网络集群服务的可信数据挖掘技术就尤为显得重要了。

2. 数据挖掘综述

数据挖掘DM(DataM ining)实质上是知识发现技术在数据库领域中的应用，它是指从大量的、不完全的、有噪声的、模糊的、随机的数据中，提取隐含在其中的、人们事先不知道的、但又是潜在有用的信息和知识的过程，[3]目前数据挖掘的主要研究内容包括基础理论、发现算法、数据仓库、可视化技术、定性定量互换模型、知识表示方法、发现知识的维护和再利用、半结构化和非结构化数据中的知识发现以及网上数据挖掘等。逻辑上，数据挖掘是可信数据以及存储在挖掘系统知识库中的知识和规则，在各种挖掘模块中处理，生成辅助模式和关系，然后进行评价，通过与分析员交互以期发现令人感兴趣的模式。有的还要加入

知识库中，以便后继的抽取和评价。(其逻辑模型见图1-1) 一般来说，它有以下几种主要数据挖掘方法：有关联规则挖掘、多层次数据汇总归纳、决策树方法等等。

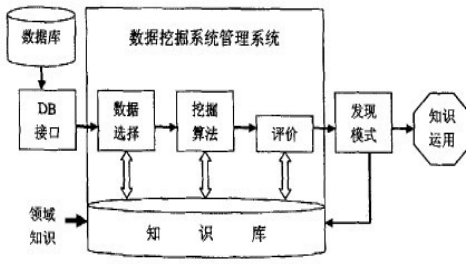

Figure 1-1. Data Mining System Logic model
图 1-1 数据挖掘的逻辑模型

3. 基于云计算集群服务的数据挖掘

3.1 问题的提出

云技术作为新兴技术现如今已有相当大的应用规模，最简单的云计算技术在网络服务中已经随处可见，例如搜寻引擎、网络信箱等，使用者只要输入简单指令即能得到大量信息。未来如手机、GPS 等行动装置都可以透过云计算技术，发展出更多的应用服务。云计算一个主要特点就是具有相当大的规模，目前 Google 云计算已经拥 100 多万台服务器， Amazon、IBM、微软、Yahoo 等的“云”均拥有几十万台服务器。企业私有云一般拥有数百上千台服务器。

基于这庞大的云计算服务集群的数据挖掘技术是一个多学科的交叉领域。力求以非寻常的方法获取蕴藏在大量看似虚拟的世界中的数据（以下都简称云数据），结合云计算服务的特点，我们提出一整套包括云数据获取、数据预处理、数据表示、数据模糊化、数据编码到算法及算法仿真的数据挖掘技术。

3.2 基于云计算的数据挖掘系统

云计算服务在这短短的几年的商业化过程中，形成了数以万计的云服务集群，这些云服务集群有一个基于服务器访问日志理念的系统，暂且称之为云服务集群日志系统，如图3-1所示，由它来监视、分析云服务数据，这就需要分析用户数据，

建立一个基于服务站点的云服务数据挖掘系统，这有助于把一般用户看来视为虚拟的，看不着摸不到的” 云”数据具体化，这个系统的建立有助于用户数据的查询、分析和操作规则化，针对这一系统的描述，建立基于云数据挖掘系统的结构图如图3-2所示规则。

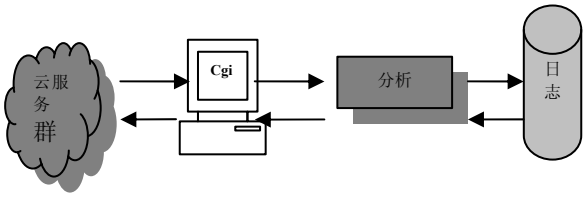

Figure 3-1. Data Mining Based on Clouds Computing Log Chart
图3-1 云服务集群日志结构实图

3.3 数据预处理

数据预处理就是对选择的云数据进行增强处理的过程。这种增强处理有时包含了根据一个或多个字段产生新的数据项，有时意味着用一个信息量更大的字段替代若干个字段。应该说明的是，输入字段的数目

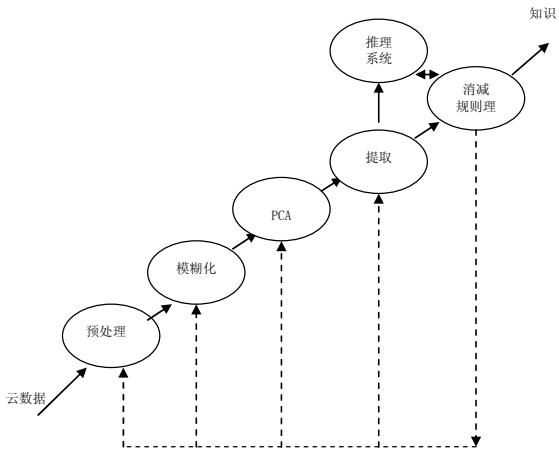

Figure 3-2. Data Mining Based on Clouds Computing System structure
图3-2 基于云服务的数据挖掘系统的结构图

不应该是提供给数据挖掘算法信息量的量度。有些数据可能是冗余数据，也就是说有些属性只不过是相同事实的不同度量方式而已。对神经网络数据挖掘来说，还需要将数据转化成一种能够被神经网络数据挖掘算法接受的形式。

3.4 数据模糊化

对输入变量进行数学处理是一个必要的工作。模糊语言使用更接近人的语言变量和语言值，所挖掘出的知识能更轻易被管理者接受，并且就目前的研究而

978-1-935068-09-9 © 2010 SciRes.

言，模糊逻辑与神经网络能够很好的结合在一起，因此本文需要对输入变量进行模糊化处理[4]。

当输入变量是连续型的取值时，需要对其进行模糊化处理。例如，对于输入变量“Hell World”，对其用“大”、“中”、“小”三个语言值来描述，每一个语言值的隶属函数都用三角形函数来表示，如图3-3所示。

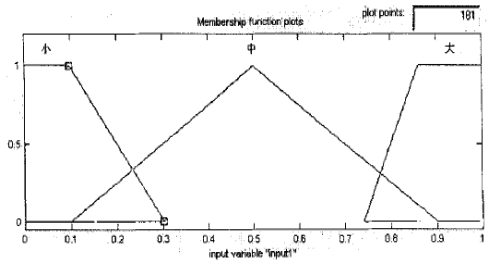

Figure 3-3. Membership function of fuzzy variable map
图3-3 模糊变量隶属函数图

这样，每一个连续型的输入变量都可以用模糊函数的隶属函数值来表示如图3-4所示。:

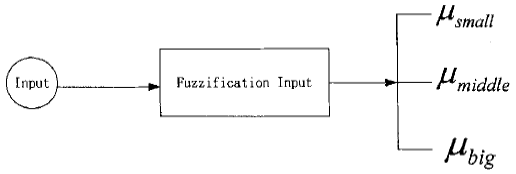

Figure 3-4. Data Mining Based on Clouds Computing Fuzzy
图3-4 输入云数据的模糊化

如果是离散型的输入变量，就将其进行0和1的编码。编码技术在此不再赘述。

3.5 主元分析(PCA)消减变量

主元分析(PCA)又叫主成分分析或主元素分析，它是多元统计过程控制方法最重要的数学工具[5]。它的主要思想是研究如何将多指标的问题转化为较少的综合指标的一种重要方法，即利用变量之间的线性相关关系对多维信息进行统计压缩，将高维空间的问题转化到低维空间去处理，使问题变得比较简单、直观，而且这些较少的综合指标之间互不相关，又能描述多维空间的绝大部分信息。

在PCA的实际应用中，可以使用总体或样本的协方差阵、离差阵或者相关阵，孟得顺等人已经证明使用样本的协方差阵和离差阵可以获得相同的结果，而协方差阵与相关阵得出的结果不同，具体使用哪个阵做主元分析取决于样本的数据信息。并且在数值分类中，由于分类性状的量纲不同，因而在度量时往往导致数量级上的差异，[6]如果数量级差异很大，用原始数据进行分类时，往往把一些重要的、但数量级很小的性状淹没，为了克服这一点，最好把数据进行模糊化处理，即把数值变到0-1之间，如极差标准化。这样做，并未把各变量(分类性状)的变异规一化，比较符合形神兼备的要求，然后按协方差阵进行主成分聚类，般能使分类结果和实际情况有较客观的符合并获得较为理想的分析结果。

本文在数据预处理阶段已经对数据进行了标准化处理并模糊化，因此拟用样本的协方差阵进行主元分析[7]。假设样本数据矩阵为X(mxn),m代表测量的采样次数，n代表测量的变量个数。主元分析方法的数据压缩过程实质上是数据矩阵X的协方差矩阵的谱分解过程

$$COV(X)=\frac{Pa}{m-1} \qquad (3.1)$$

$$COV(X)P_i=\lambda_i P_i \qquad (3.2)$$

其中，COV(X)是X的协方差阵，凡是按降序排列的协方差阵的特征值，P，特征向量。主元空间的信息抽取实质上是选择几个有代表性的主元，解释数据中的大部分变化，数学表达式如下:

$X= t_1p_1+t_2p_2+\ldots+t_kp_k,+E$ 式中，$t_1(i=1,2,\ldots,k)$是系统主元，也称为得分向量，提取采样数据间关联信息;$P_i(i=1,2,\ldots,k)$是主元特征向量，也称为载荷向量，提取变量间关联信息:E是残差矩阵，提取噪声和模型误差信息。各个得分向量之间是正交的，即对任何i和j当$i\neq j$时满足$t_it_j=0$。各个负荷向量之间也是互相正交的，同时每个负荷向量的长度都是1，即

$$P_iP_j=0 \quad (i\neq j) \qquad (3.3)$$

$$P_iP_j=0 \quad (i==j) \qquad (3.4)$$

上式两边同时乘以P_i，就可以得到$t_i=XP_i$，向量t_i的长度反映了数据矩阵X在P_i方向上的覆盖程度。它的长度越大，X在P_i方向上的覆盖程度或者变化范围也就越大。那么负荷向量P_i代表X变化的最大方向，P_i与P_j垂直并代表X变化的第一大方向，P_z将代表变化最小的方向。当矩阵X中的变量间存在一定程度的线性相关时，X的变化主要体现在最前面的几个负荷向量方

向上，X的最后几个投影比较小的负荷向量，可以写成残差矩阵E,主要由噪声引起，往往可以忽略，起到减少噪声影响的效果，不会引起数据中有用信息的明显损失。在信息抽取过程中，合理确定主元的个数非常重要，主元选取多则模型相对精确，但增加了分析与诊断的复杂性，无法有效清除噪声，主元选取过少，则不能充分提取原始数据空间的信息，使分析与诊断的误差率增加。在实际检验中，可采用交叉检验法来确定最优主元个数，用来检验所建立的主元回归模型。通过保留不同数据的主元，建立若干主元回归模型，然后在检验数据上测试这些模型，并从中选取在检验数据中测试误差最小的那个主元回归模型。或者也可用主元贡献率累积和百分比(CPV)的方法，先计算各主元贡献率，选择累积和百分比大于某个百分数的主元个数。

3.6 系统规则提取

从云服务群集获得数据经过模糊化和编码后，数据很难得到很高的精度，且是其获得的知识隐含在云服务群集的一系列连接和权值中，处理过程无法为用户所理解，模型对于用户来说是一个黑箱[8]。由于缺乏透明性，在数据挖掘和决策支持领域，以及在安全性能要求很高的关键应用方面，往往被认为是不可靠和难于理解的，应用受到一定限制。因此，有必要建立一个解释机制，用规则取代权值矩阵，为决策支持类应用提供完整的决策说明，为关键应用提供结果的可信度和质量检测手段，并为基于符号和基于连接的两种AI技术提供一种有效的集成方法。

关于数据规则提取的算法颇多，任选合适则可，这些算法基本上是基于数学公式推导，计算过程较为繁琐复杂。针对云计算服务的特点，本文提出了一种基于结构分解的规则提取算法，它以隐藏结点和输出结点为单位将网络分解为若干单层网络的集合，对每一子网搜索和提取规则，最后对这些规则进行组合以描述整个网络的特性。所提取的规则有：

if X_i is Y_i，and X_z is Y_z …and X_o is Y_o then C,weight；其中X_i代表一输入变量，Y_i代表该输入变量的模糊隶属函数值或者该输入变量的编码，C代表类别，weight表示该规则的权重[9]。规则权重的大小表示此规则在规则库中的强弱程度，体现了此规则描述现实的能力。

3.7 使用遗传算法消减规则

遗传算法(Genetica lgorithms,简称GA)是一种基于自然选择和基因遗传学原理的搜索算法，它将“优胜劣汰，适者生存”的生物进化原理引入待优化参数形成的编码串群体中，按照一定的适值函数及一系列遗传操作对各个体进行筛选，从而使适值高的个体被保存下来，组成新的群体[10-11]。新群体包含上一代的大量信息，并且引入了新的优于上一代的个体。这样周而复始，群体中各个适应度不断提高，直至满足一定的极限条件。此时，群体中适值最高的个体即为待优化参数的最优解。使用遗传算法进行规则消减过滤，就是要从先前提取出的规则里面通过遗传迭代找出一部分最能够代表正确分类的那些规则，删去有可能不确信的规则，力求用最精简和最精确的语言来正确的对目标数据库进行分类。适值函数可用下式评估：

$$f(s) = NC\ P(s)$$

其中MCP为能被S正确分类的数据个数，由模糊推理系统计算。首先随机产生n个长度为二的染色体种群，然后由模糊推理系统评估侮一个染色体字符串的适应度，并挑选出适应值较高的前50%的染色体作为父群体进行繁衍[12]。本文所使用的交叉算法分为两个步骤。第一步是决定两两交叉的对象，山计算机随机在父染色体群体中抽取两两匹配的对象。第二步是决定交换点及交换规则，由计算机随机选取。例如计算机随机选取代表规则权重的两个子串S,和Sz进行交叉，并随机选择两个染色体的第一位和第三位进行代码交换:

1001…01　　0011…01

0110…11　　1100…11

尽管复制和交叉操作很重要，但不能保证不会遗漏一些重要的遗传信息。在人工遗传系统中，变异用来防止这种不可弥补的遗漏。变异就是某个字符串某一位的值偶然的随机改变，即在某些特定的位置上把1变为0，或反之[13]。当它有节制的和交叉一起使用时，它就是一种防止过度成熟而丢失重要概念的保险策略。例如随机产生一个种群;

S_1　01101

S_2　11001

S_3　00101

S_4　11100

这样很明显看到，无论怎样交叉，位置4上都不可得到有1的位串。若优化结果值要求此位上位1，显然交叉不能搜索到最优结果，因此需要变异来进行必要的转变。变异操作可以起到恢复位串字符多样性的作用，

978-1-935068-09-9 © 2010 SciRes.

并能适当的提高遗传算法的搜索效率。根据经验的研

Table 1. Comparison of genetic algorithm rule base
表 1 遗传算法规则库比较

规则提取	Iris Species		Pima Indian		Credit Approval		Heart Disease	
	规则数	前件数	规则数	前件数	规则数	前件数	规则数	前件数
原始规则库	21	21	40	40	46	46	65	65
消减后规则库	7	7	15	15	17	17	22	22

究，为了取得较好的结果，本文设定变异率为0.001，即变异的频率为每一个千位的传送中，只变异一位。使用该算法，对消减前后规则库样数对比（表1）可以看出，基于云计算的网络服务采用本文提出的规则提取方法，所提取的规则数明显少于那些一般算法规则的规则数。对于用遗传算法过滤后的规则库，知识表达更加准确精练。

4. 结束语

基于云计算的集群服务模式是近几年来主要的商业运行模式，它类似但又区分于一般的网格计算、分布式计算、并行计算，再加上数据及数据挖掘任务和数据挖掘方法的多样性，这就给研究基于云计算集群服务模式方面的数据挖掘提出了许多颇具挑战性的课题。本文结合云计算集群服务模的特点，给出了一个基于云计算服务的数据挖掘系统和一整套研究数据挖掘的流程，详细介绍了原始数据的装备、填入数据的模糊化及数据的消减、规则的提取、算法描述等等，最后比较了算法前后规则库中数据的变化，体现了系统对基于云计算服务模式下数据描述性的普遍适应性和规则研究表达的准确和精炼性。

4 致谢

本论文的完成是在几位老师的大量指导和帮忙下完成的。特别感谢章荣丽老师的帮忙和参与，为论文研究的方向、框架、论证和实施做出了相当的工作，同时感谢我的同事，感谢他们的建议和支持，他们对待学问研究的刻苦专研、勇于创新、孜孜不倦的态度和精神，使我受益匪浅、且颇受感动。

References (参考文献)

[1] http://www.cloudcomputing-china.cn/
[2] http://www.chinabyte.com/188/8045188.shtml
[3] J.P ei,J .H an,B .M ortazavi-Asl,a nd1 1.Z hu”MiningA ccessP atterns Ef fi ci ently fromW eb Logs (PUP)”，Proc.2 000 Pacific-AsiaC onf.o n
[4] Ramakrishnan Srikant，Yinghui Yang. Mining Log Improve Site
[5] Or ga ni zation
[6] Mulvann MD, Discovering Internet marketing Intelligence through on lin e A nalyticalw ebu sagem ining.L JlA CMS IGMODR ecord,19 98,27 (4)
[7] JianP ei,J iaweiI lan.B ehxadM ortazavi-asd,a ndH ua Zhu. MiningA ccess Pa tt er nsE fficiently from Web Logs.
[8] Pi a tet sky-ShapiroG .D ataM ininga ndK nowledgeD iscoveryi nB usiness Databases. ISMIS'96,56-57
[9] M in ingA ssociationR ulesB etween Setso fI temsi nL argeD atabases.Proc. ACM SIGMODi nt'l c onf.M anagemento fD ata,W ashington.D C,M ay1 993,20 7-216
[10] Yusuke Ohura, Katsumi Takahashi, Iko Pramudiono, Masaru Kitsuregawa. Ex pr ei mentso nQ ueryE xpansionf orI nternetY ellowP ageS ervicesU sing We b Lo g Ming
[11] TANG, Fu-Hua Yang, Yang Lu. The basic method of data mining and its differences with the expert system. Computer applications, Vo119, N o3, 1 999M.-S.Ch en, et al
唐常杰、杨富华、杨璐.数据挖掘的基本方法及其与专家系统的差异.计算机应用，Vo119,N o3,1 999M.-S.Ch en,et al .
[12] Retired intellectuals, Deng Su, Zhang Weiming. Data Mining and Knowledge Discovery. Calculation sector Monde, 24 version of the topics, 1997.6.30
陈文伟、邓苏、张维明.数据挖掘与知识发现综述.计算界世界报，24期专题版，1997.6.30
[13] Anupam Joshi. On Mining Web Access Logs.
[14] R. Ag rawal,e ta l.D atabaseM ining:A Performance Perspective. IEEE Tansactions on knowledge and data engineering, 1993, 5(6), 914-925