doi:10.3969/j.issn.1000-436x.2013.10.015

大数据典型相关分析的云模型方法

杨静，李文平，张健沛
（哈尔滨工程大学 计算机科学与技术学院，黑龙江 哈尔滨 150001）

摘 要：针对传统大数据典型相关分析（CCA, canonical correlation analysis）方法的高复杂度在面临大数据 PB 级数据规模时不再适应的现状，提出了一种基于云模型的大数据 CCA 方法。该方法在云计算架构的基础上，通过云运算将各端点云合并为中心云，并据此产生中心云滴，以中心云滴作为大数据的不确定性复原小样本，在其上施以 CCA 运算，中心云滴的较小数据量提高了运算效率。在真实数据集上的实验结果验证了该方法的有效性。
关键词：大数据；典型相关分析；云模型；云运算；云计算
中图分类号：TP391 **文献标识码：**B **文章编号：**1000-436X(2013)10-0121-14

Canonical correlation analysis of big data based on cloud model

YANG Jing, LI Wen-ping, ZHANG Jian-pei
(College of Computer Science and Technology, Harbin Engineering Univers ty, Harbin 150001, China)

Abstract: The complexity of traditional CCA methods is too high meet the requirements to analyze big data due to their huge scale which is reaching the level of peta-byte. A novel approach to CCA was proposed to mine the big data by introducing the cloud model which is a brand-nowel theory about the uncertainty artificial intelligence. A distributed architecture based on cloud computing was established. All of the clouds distributing on the nodes of the distributed architecture were combined to a center cloud via cloud operation (whe cloud is a synopsis of data and which is a concept coming from the cloud theory). A type of virtual sample of data called cloud drops created based on the center cloud. Finally the computing of CCA was imposed on the cloud drops. The CCA was impose on the cloud drops with less volume, which improves the efficiency. Experimental results on real data sets indicate the effectiveness of this method.
Key words: big data; canonical correlation analysis (CCA); cloud model; cloud operation; cloud computing

1 引言

自 2008 年 9 月《Nature》杂志推出名为“大数据”（big data）的封面专栏[1]以来，产业界和学术界便掀起了大数据研究热潮。数据量巨大是大数据的首要特性，通常认为 PB 级别及其以上的数据称为“大数据”。大数据还具有稀疏价值特性，即大数据所携带的信息在刻画某特定知识方面是冗余的。这些特性为大数据挖掘带来了巨大的挑战。

大数据典型相关分析（CCA, canonical correlation analysis）是大数据研究的重要内容之一，它不仅有助于揭示大数据间的相关关系，而且可提取蕴含于大数据中的低维特征。大数据 CCA 可用于大数据特征融合[2]、机器学习[3]、数据降维[4]、数据流挖掘[5]等领域。因此大数据 CCA 具有重要的意义。

大数据 CCA 研究极具挑战性，其困难不仅源于 CCA 本身具有的高复杂度，而且也来自大

收稿日期：2013-04-21；修回日期：2013-07-30
基金项目：国家自然科学基金资助项目(61370083, 61073043, 61073041)；高等学校博士学科点专项科研基金资助项目(20112304110011, 20122304110012)；黑龙江省自然科学基金资助项目（F200901）；哈尔滨市科技创新人才研究专项基金（优秀学科带头人）资助项目(2011RFXXG015)
Foundation Items: The National Natural Science Foundation of China(61370083, 61073043, 61073041); The Research Fund for the Doctoral Program of Higher Education of China(20112304110011, 20122304110012); The Natural Science Foundation of Heilongjiang Province (F200901); The Harbin Special Funds for Technological Innovation Research(2011RFXXG015)

数据巨大规模以及稀疏价值等特性。面向传统数据的 CCA 方法的高空间复杂度在面临大数据 PB 级规模时已不再适应。针对此问题，本文拟研究一种基于云模型的大数据 CCA 方法，期望该方法能克服大数据巨大规模所带来的高复杂度等困难。

云理论是一种实现定量数据和定性概念之间相互转换的不确定性人工智能方法，最早由我国学者李德毅院士提出。云的具体实现称为云模型。云模型在信任评估[6,7]、时间序列挖掘[8]以及图像分割[9]等广泛领域得到了成功应用。然而，将云模型与 CCA 结合，以用于大数据研究还鲜有学者涉足，本研究拟在此方面展开初探工作。

本文首先根据逆向云发生器生成各云端的数据概要；其次将数据概要发送至中心云端，利用云运算操作产生中心云数字特征；最后根据中心云数字特征，利用正向云发生器产生中心云滴，在中心云滴上施加 CCA 操作。中心云数字特征刻画了各云端中数据的语言值，据此产生的中心云滴是原来大数据的不确定性复原小样本。中心云滴在概念粒度上携带了原始数据的重要信息，从这个意义上来说，研究中心云滴不是在原始数据上直接计算，是探讨大数据挖掘的一个良好视角；此外，中心云滴的小样本特性为 CCA 赢得了效率。

2 基础知识回顾

2.1 CCA

CCA 是研究 2 个随机向量之间相关性的一种常用多元统计方法[10]。给定 p 维随机向量 $\boldsymbol{X}$ 和 q 维随机向量 $\boldsymbol{Y}$，$p \leq q$，CCA 的目标是寻找投影向量 $\boldsymbol{a}_k$ 和 $\boldsymbol{b}_k$，使得在方差 $\mathrm{var}(\boldsymbol{a}_k^{\mathrm{T}}\boldsymbol{X}) = \mathrm{var}(\boldsymbol{b}_k^{\mathrm{T}}\boldsymbol{Y}) = 1$ 的约束下，Pearson 相关系数

$$r(\boldsymbol{a}_k^{\mathrm{T}}\boldsymbol{X},\ \boldsymbol{ß}_k^{\mathrm{T}}\boldsymbol{Y}) = \frac{\boldsymbol{a}_k^{\mathrm{T}}\boldsymbol{C}_{xy}\boldsymbol{ß}_k}{\sqrt{\left(\boldsymbol{a}_k^{\mathrm{T}}\boldsymbol{C}_{xx}\boldsymbol{a}_k\right)\times\left(\boldsymbol{ß}_k^{\mathrm{T}}\boldsymbol{C}_{yy}\boldsymbol{ß}_k\right)}} \quad (1)$$

达到最大值。其中，$\boldsymbol{C}_{xy} = \boldsymbol{C}_{yx}^{\mathrm{T}} = \boldsymbol{X}\boldsymbol{Y}^{\mathrm{T}}$ 为 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 之间的互协方差矩阵，而 $\boldsymbol{C}_{xx} = \boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}$ 和 $\boldsymbol{C}_{yy} = \boldsymbol{Y}\boldsymbol{Y}^{\mathrm{T}}$ 分别为 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 的自协方差矩阵。称 $\boldsymbol{a}_k^{\mathrm{T}}\boldsymbol{X}$ 和 $\boldsymbol{b}_k^{\mathrm{T}}\boldsymbol{Y}$ 为 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 的第 k 对典型相关变量，其相关系数称为第 k 个典型相关系数。

CCA 实质是一个最优化问题。以第一对典型变量为例（省略 $\boldsymbol{a}_1$ 和 $\boldsymbol{ß}_1$ 下标），即求

$$\max_{\boldsymbol{a}\in\mathrm{R}^{p\times1},\,\boldsymbol{ß}\in\mathrm{R}^{q\times1}} \boldsymbol{a}^{\mathrm{T}}\boldsymbol{C}_{xy}\boldsymbol{ß},\ \text{s.t.}\ \boldsymbol{a}^{\mathrm{T}}\boldsymbol{C}_{xx}\boldsymbol{a}=1,\ \boldsymbol{ß}^{\mathrm{T}}\boldsymbol{C}_{yy}\boldsymbol{ß}=1 \quad (2)$$

其中，s.t. 表示约束条件，R 为实数域。用拉格朗日（Lagrange）乘子法求解式(2)有

$$\boldsymbol{ß} = \frac{1}{l}\boldsymbol{C}_{yy}^{-1}\boldsymbol{C}_{yx}\boldsymbol{a} \quad (3a)$$

$$\boldsymbol{C}_{xy}\boldsymbol{C}_{yy}^{-1}\boldsymbol{C}_{yx}\boldsymbol{a} = l^{2}\boldsymbol{C}_{xx}\boldsymbol{a} \quad (3b)$$

式(3b)是广义特征值问题，由此解出 l 和 $\boldsymbol{a}$，代入式(3a)可得 $\boldsymbol{b}$。l 即为所求典型相关系数。CCA 有多种解法，如基于 SVD 的方法等，具体可参阅文献[11,12]。

2.2 云和云模型

设 U 为定量论域，C 为其上的定性概念，若 $\forall x \in U$ 是 C 的随机实现，且 x 对 C 的确定度 $m(x) \in [0,1]$ 是有稳定倾向的随机数。

$$m: U \to [0,1],\quad x \to m(x)$$

则 x 在 U 上的分布称为云（cloud），而 x 称为云滴（cloud drop）[13]。云理论用期望 Ex、熵 En 和超熵 He 3 个数字特征来表征概念的整体定量特性。在不至混淆时，也将云的 3 个数字特征构成的三元组 (Ex, En, He) 称为云。

云模型是云的具体实现。由云数字特征产生云滴的实现称为正向云发生器，而由云滴群得到云数字特征的实现称为逆向云发生器。由于正态分布的普适性，建立在其上的正态云是各种云模型中最重要的一种。期望曲线是云理论研究数据集在空间中随机分布统计规律的重要方法，一般方程为

$$y = \exp(-(x-Ex)^2/2(En)^2) \quad (4)$$

云运算是云理论中用语言值进行计算和推理的重要基础。给定 2 个一维云 $C_1(Ex_1, En_1, He_1)$ 和 $C_2(Ex_2, En_2, He_2)$，则 C_1 加 C_2 之和 $C(Ex, En, He)$ 可以定义为

$$\begin{aligned} Ex &= Ex_1 + Ex_2 \\ En &= \sqrt{(En_1)^2 + (En_2)^2} \\ He &= \sqrt{(He_1)^2 + (He_2)^2} \end{aligned} \quad (5)$$

需要补充的是，“云”一词有趣地同时光顾了云计算和云理论，为了不至于混淆，本文所述云端皆指云计算平台中的分布式节点或机群，而其他关于云的词汇，特指云理论中的概念。此外，应将云

运算和云计算区别开来。云运算是云理论中对云进行操作的规则，属于不确定性人工智能范畴；而云计算是一种计算范式，强调计算资源的有效利用和整合，与云运算截然不同。

3　相关工作

人类在科研和工程实践项目中收集的大量数据多数具有大数据特性，但将大数据抽象出来作为一门独立科学进行研究还是最近的事[14]。在生物信息学等领域，Benjamin 等人深入研究了在系统神经生物学领域担当重要角色的生理电大数据压缩及存储等问题[15]；Aronova 等人将生物学研究中收集的数据视为大数据，从大科学（big science）视角挖掘这类数据蕴含的重要知识[16]；Werner 则更进一步，从方法论角度分析了如何应对大数据生物学带来的挑战[17]。

在数据挖掘等领域，Alfredo 等人从数据仓库和 OLAP 等视角分析了多维大数据研究存在的问题以及研究趋势[18]；Steven 等人研究了大数据挖掘中的在线特征选择问题[19]；Simon 等人基于模糊查找词典（fuzzy find dictionary）研究了一种面向数据流大数据的数据流聚类方法[20]；John 研究了大数据上的并行学习问题[21]。

在面向大数据的程序开发和存储等方面，Thomas 等人探讨了如何在大数据上构建程序实现问题[22]；Yu 等人提出了一种可扩展的用于大数据分析的分布式系统[23]；Kyuseok 以及 Jens 等人同时探讨了 MapReduce 架构在大数据分析中的应用[24,25]；Divyakant 等人分析了大数据及云计算现状和研究挑战[26]；Huiqi 等人研究了在云平台上进行可视聚类的一种方法体系[27]。此外，也有学者开始涉足大数据安全方面的研究，如 Colin 等人探讨了大数据中存在的安全问题及解决策略[28]。

大数据研究还刚刚起步，尽管有学者探讨了基于云计算平台的大数据存储方法，但未发现关于大数据 CCA 的研究报告，也未发现在此方面基于云理论的研究方法，期望本研究能对此做出些许初探性工作。

4　大数据 CCA 方法

本节重点研究基于云模型的大数据 CCA 方法（BDCCA, big data CCA）。首先阐述面向大数据的云架构，其次重点探讨端点云的生成方法，再次研究端点云的合并技术。下文约定运算符 $<\bullet,\bullet>$ 为欧氏内积，而 $\otimes$ 为 Hadamard 积。

4.1　面向大数据的分布式云架构

就容量而言，PB 级数据量被认为是大数据的显著特性，这一特性使得大数据一般通过机群等分布式方式存储。迄今为止，云平台是大数据存储的理想载体。本研究假设大数据以分布式方式存储在云端。图 1 刻画了所提出的由若干个云端构成的大数据分布式云架构。

此云架构从功能上分 4 层：1）顶层为数据存储层，其中，第 i 个云端存储第 i 段数据 $Data_i$；2）第 2 层为多维逆向云发生器（MBCG, multidimensional backward cloud generator）层，其核心任务在于由原始数据产生各云端的云，即端点云；3）第 3 层为中心云端（center node），该层主要进行云合并运算，并用于产生和存储中心云滴；4）第 4 层为应用层（applications），基于中心云滴，在此层可进行 CCA 等挖掘或分析任务。

在大数据分布式云架构中：1) 根据多维逆向云发生器 MBCG，由第 i 个云端中的数据 $Data_i$ 产生端点云 $C_i\ (Ex_i, En_i, He_i)$，简记为 C_i；2) 将 C_i 传送至中心云端的云收集器（collector）；3) 将云收集器中的云传送至多维云合并节点（MCC, multidimensional cloud combiner）；4) 根据多维云合并运算，将所有云 C_i 合并为中心云 $C(Ex, En, He)$，简记为 C；5) 将中心云 C 传送至多维正向云发生器 (MFCG, multi- dimensional forward cloud generator) 节点；6)根据 MFCG，由中心云 C 产生中心云滴；7)应用层中 CCA 等任务到中心云端获取中心云滴，并据此进行相应的挖掘任务。

此云计算架构用于处理大数据是合适的。1)各云端向中心云端仅传送数据概要，即由云数字特征构成的三元组，如此小的数据量传送是快速的；2) 由中心云产生的中心云滴群规模往往较小，这有助于提高 CCA 的运算速度。

4.2　BDCCA 执行流程

BDCCA 的基本思路在于：1)在各云端利用逆向云发生器根据当前云端中数据并行生成云（即云数字特征）；2)将各端点云发送至中心云端，利用多维云合并操作，在中心云端产生中心云；3) 根据中心云，利用正向云发生器产生中心云滴；4)在中心云滴上施加 CCA 操作。图 2 描述了其执行流程。

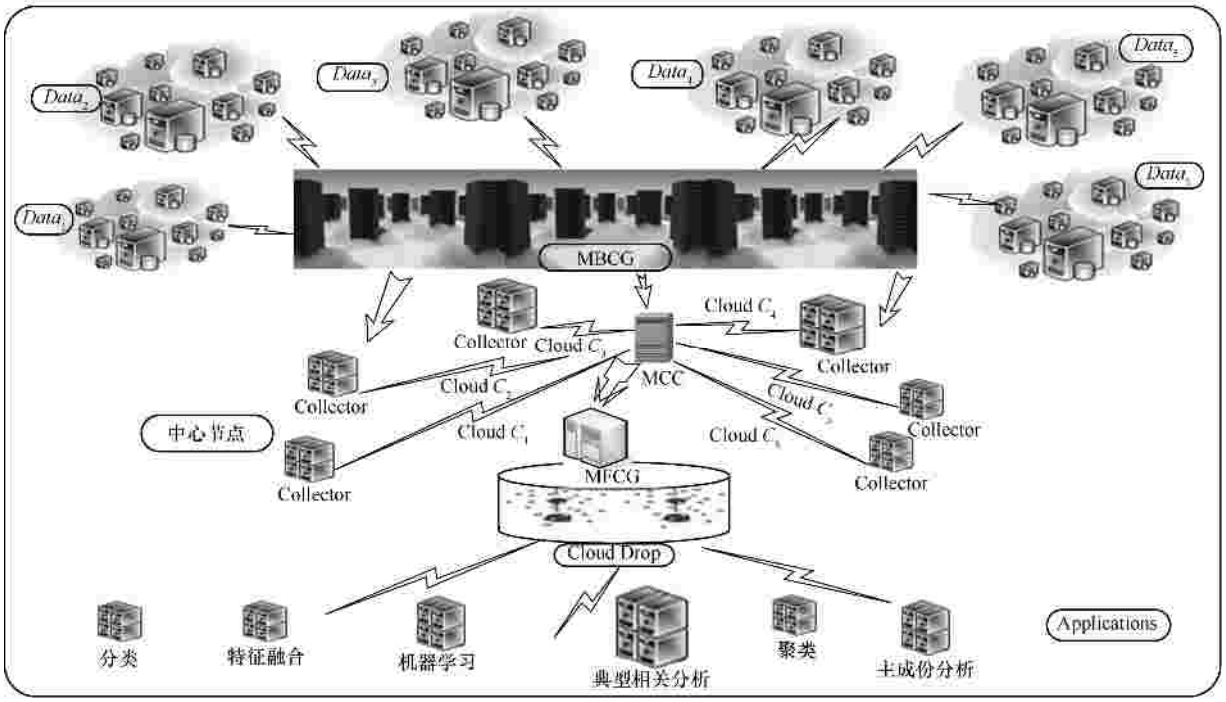


图 1 大数据分布式云架构

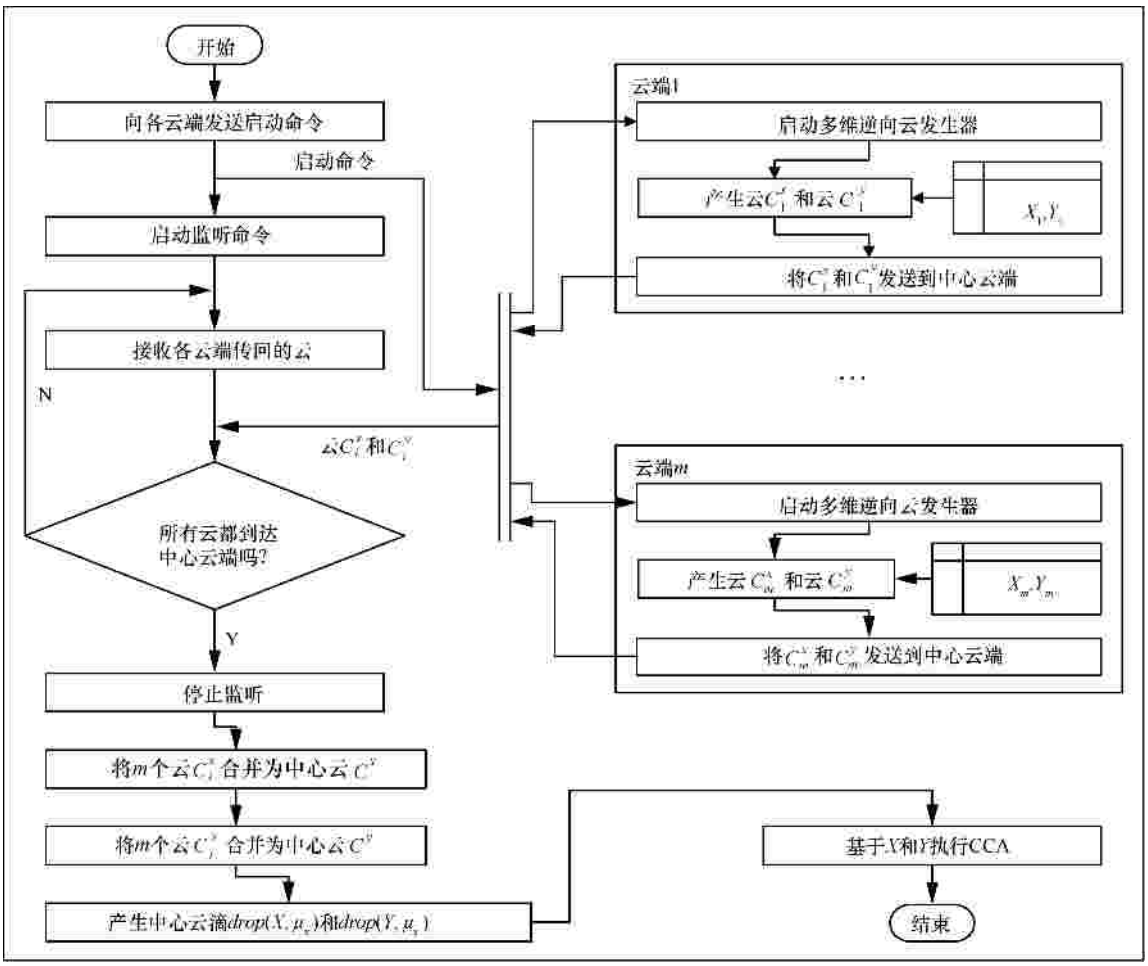


图 2 BDCCA 执行流程

数据在每个云端分为 $\boldsymbol{X}_i$ 和 $\boldsymbol{Y}_i$ 两部分，其中，$\boldsymbol{X}_i \in \mathrm{R}^{p\times n_i}$ 和 $\boldsymbol{Y}_i \in \mathrm{R}^{q\times n_i}$，$n_i$ 为第 i 个云端中的样本数目，p 为 $\boldsymbol{X}_i$ 的维数，q 为 $\boldsymbol{Y}_i$ 的维数。特别地，同类数据的维数在所有云端都一致，而样本数目可以不同。此外，云端个数 m、各云端标识符 N_i、云重要度向量 $\boldsymbol{h}=(h_1,h_2,\mathrm{L},h_m)^{\mathrm{T}}$ 以及中心云滴数目 w 等需预先设定。流程执行结束后，输出典型相关系数向量 $\boldsymbol{r}$ 以及对应典型相关向量为列的矩阵 $\boldsymbol{U}$、$\boldsymbol{V}$。基于式(3)，可通过特征分解或 SVD 等方法求解 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 的典型相关系数和典型相关变量，具体可参阅文献[11]。本文将采用文献[30]的多维正向正态云发生器产生中心云滴群 $drop(\boldsymbol{X},\boldsymbol{m}_x)$ 和 $drop(\boldsymbol{Y},\boldsymbol{m}_y)$。限于篇幅，此两点不再赘述。

图 2 所示流程中，产生各端点云以及在中心云端进行云合并是关键，后文将分别详述这两点，一方面后文将对多维逆向云发生器进行改进，使之适宜于在大数据环境下产生各端点云；另一方面将提出一种一次合并多个多维云的方法，以提高大数据环境下云合并运算的效率。

4.3　端点云生成

所谓端点云的生成，是指根据逆向云发生器，由云端中数据产生云的过程。本文采用无确定度的多维逆向正态云发生器[30]作为端点云的生成模型。

尽管已将大数据存储于分布式云架构各云端（如图 1 所示），但是由于大数据的巨大容量特性，在每个云端所存储的数据量往往还较大，现存多维逆向正态云发生器不再满足大数据环境下计算效率的要求，对之加以改进是必要的。

为了提高多维逆向正态云发生器在大数据环境下产生云的效率，本文基于随机采样法，采用启发式云生成策略，将多维逆向正态云发生器拓展到大数据情形。

4.3.1　大数据随机采样

本文借鉴随机子空间法[29]思想，在各云端进行大数据随机采样。设各云端将大数据分为若干块，首先对每块按照相同划分方式将其分割成 s 个子块；然后将所有块中相同位置的子块转换成列向量并进行组合，形成一个子块集，如图 3 所示。

基于划分的数据块，在每个子块集上执行随机采样。对第 i 个子块集 T_i，根据随机子空间法思想，随机产生 r^* 维索引向量 $\boldsymbol{I}_i=\{j_1,j_2,\mathrm{L},j_{r^*}\}$，$r^*<r$，$r$ 为子块集大小。将所有子块集中产生的索引向量按下标升序组合为 $\boldsymbol{I}=\{I_1,I_2,\mathrm{L},I_s\}$。对每个云端数据 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 分别执行上述操作。最后在与索引向量对应数据上执行 CCA 操作。

4.3.2　云生成的启发式策略

在每个云端，云的产生采用启发式策略。其基本思想是，在每个云端迭代地进行若干次不重复随机采样，将每次迭代时抽取的样本加入之前的样本中，每次迭代后进行云更新，若第 i 次迭代后所生成的云 C_i 与迭代前的云 C_{i-1} 之差 D_{C_i} 小于给定阈值或迭代次数超过预设阈值，则迭代终止。迭代过程中，若当前迭代的云差异 D_{C_i} 正向偏离前一次迭代的云差异 $D_{C_{i-1}}$，即 $D_{C_i}-D_{C_{i-1}}>d$，则下一次迭代时将加大随机采样的样本容量；反之若 D_{C_i} 负向偏离 $D_{C_{i-1}}$，即 $D_{C_i}-D_{C_{i-1}}\le d$，则下一次迭代时将减小随机采样的样本容量。其中，$d\ge 0$，d 为常量。

此策略的 2 个关键问题在于，其一每次迭代后云的更新；其二相邻两次更新所生成云之间差异的刻画或度量。

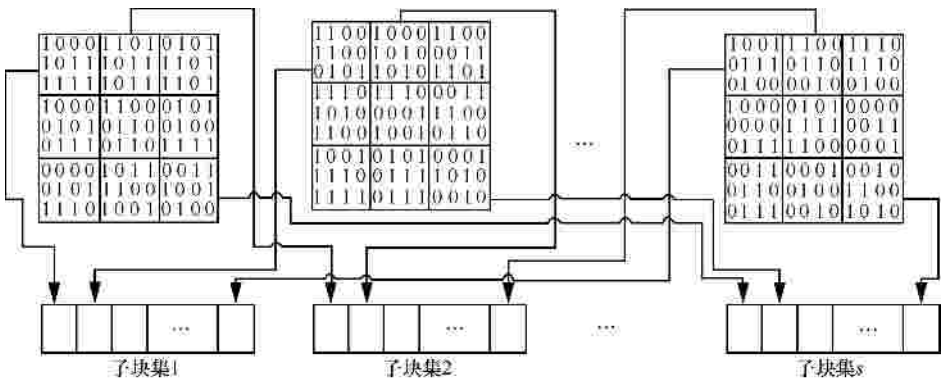


图 3　数据子块划分

4.3.3 云的部分增量式更新

每次迭代后的云更新是云的启发式生成策略需要解决的首要问题。云更新即是云期望 $Ex=\frac{1}{n}\sum_{k=1}^{n}x_k$、熵 $En=\sqrt{\pi/2}A$ 和超熵 $He=\sqrt{S^2-En\otimes En}$ 的更新。其中

$$A=\frac{1}{n}\sum_{k=1}^{n}|x_k-\bar{x}|$$

$$S^2=\frac{1}{n-1}\sum_{k=1}^{n}(x_k-\bar{x})\otimes(x_k-\bar{x})$$

若记

$$Ex_i=\bar{x}_i=\frac{1}{n_i}\sum_{k=1}^{n_i}x_k$$

$$G_i=\sum_{k=1}^{n_i}(x_k-\bar{x}_i)\otimes(x_k-\bar{x}_i)$$

$$\delta_i=\sum_{k=1}^{n_i}|x_k-\bar{x}_i| \tag{6}$$

其中，n_i 为第 i 次迭代后的样本 $\{x_k\}_{k=1}^{n_i}$ 总容量，而 $\tilde{n}_i$ 为第 i 次迭代进行随机采样所得的样本 $\{\tilde{x}_k\}_{k=1}^{\tilde{n}_i}$ 容量，显然 $n_i=n_{i-1}+\tilde{n}_i$。云增量式更新的本质在于：①用 Ex_{i-1} 刻画 Ex_i；②根据 G_{i-1} 求解 G_i；③由 δ_{i-1} 计算 δ_i。本研究主要更新前两者，故称为部分增量式更新。

设第 $i-1$ 次迭代后所生成的云为 $C_{i-1}(Ex_{i-1},En_{i-1},He_{i-1})$，并记第 i 次迭代进行随机采样所得样本对应的云为 $C_{\tilde{i}}(\tilde{Ex}_i,\tilde{En}_i,\tilde{He}_i)$。则第 i 次迭代后所得云的期望为

$$Ex_i=\frac{n_{i-1}Ex_{i-1}+\tilde{n}_i\tilde{Ex}_i}{n_i} \tag{7}$$

这只需注意到

$$\begin{aligned}Ex_i&=\frac{1}{n_i}\sum_{k=1}^{n_i}x_k\\&=\frac{n_{i-1}}{n_i}\frac{1}{n_{i-1}}\sum_{k=1}^{n_{i-1}}x_k+\frac{\tilde{n}_i}{n_i}\frac{1}{\tilde{n}_i}\sum_{k=1}^{\tilde{n}_i}\tilde{x}_k\\&=(n_{i-1}Ex_{i-1}+\tilde{n}_i\tilde{Ex}_i)/n_i\end{aligned}$$

其中，$\tilde{Ex}_i=\frac{1}{\tilde{n}_i}\sum_{k=1}^{\tilde{n}_i}\tilde{x}_k$ 为样本 $\{\tilde{x}_k\}_{k=1}^{\tilde{n}_i}$ 的均值向量。记 $G_i=\sum_{k=1}^{n_i}(x_k-\bar{x}_i)\otimes(x_k-\bar{x}_i)=\sum_{k=1}^{n_i}(x_k-\bar{x}_i)^2$，$D_1=\sum_{k=1}^{\tilde{n}_i}(\tilde{x}_k-\bar{x}_i)^2$，$D_2=(\bar{x}_{i-1}-\bar{x}_i)^2$，并留意到 $\sum_{k=1}^{n_{i-1}}(x_k-\bar{x}_{i-1})=0$，则有

$$\begin{aligned}G_i&=\sum_{k=1}^{n_{i-1}}(x_k-\bar{x}_i)^2+\sum_{k=1}^{\tilde{n}_i}(\tilde{x}_k-\bar{x}_i)^2\\&=\sum_{k=1}^{n_{i-1}}(x_k-\bar{x}_{i-1})^2+2(\bar{x}_{i-1}-\bar{x}_i)\sum_{k=1}^{n_{i-1}}(x_k-\bar{x}_{i-1})+\\&\quad(\bar{x}_{i-1}-\bar{x}_i)^2\sum_{k=1}^{n_{i-1}}1+\sum_{k=1}^{\tilde{n}_i}(\tilde{x}_k-\bar{x}_i)^2\\&=G_{i-1}+D_1+n_{i-1}D_2\end{aligned} \tag{8}$$

可得：

$$\begin{aligned}S_i^2&=\frac{1}{n_i-1}G_i\\&=\frac{n_{i-1}-1}{n_i-1}S_{i-1}^2+\frac{D_1+n_{i-1}D_2}{n_i-1}\end{aligned} \tag{9}$$

由于绝对值缺乏良好的代数性质，因此要获得 A 的增量表达式是困难的。本研究在迭代过程中只需跟踪云期望向量 Ex_i 和中间向量 G_i 即可，而不需跟踪 δ_i 的改变量。定理 1 阐述了其理由。

定理 1 令 $D_{G_i}=G_i-G_{i-1}=D_1+n_{i-1}D_2$，$D_{\delta_i}=\delta_i-\delta_{i-1}$，$D_1=\sum_{k=1}^{\tilde{n}_i}(\tilde{x}_k-\bar{x}_i)^2$，$D_2=(\bar{x}_{i-1}-\bar{x}_i)^2$。则当 $D_{G_i}\to 0$ 时，$D_{\delta_i}\to 0$。

证明 $D_1=\sum_{k=1}^{\tilde{n}_i}(\tilde{x}_k-\bar{x}_i)^2\geq 0$，$D_2=(\bar{x}_{i-1}-\bar{x}_i)^2\geq 0$，所以当 $D_{G_i}\to 0$ 时，$D_1\to 0,D_2\to 0$，从而有

$$\begin{aligned}\delta_i&=\sum_{k=1}^{n_{i-1}}|x_k-\bar{x}_i|+\sum_{k=1}^{\tilde{n}_i}|\tilde{x}_k-\bar{x}_i|\\&\to\sum_{k=1}^{n_{i-1}}|x_k-\bar{x}_{i-1}|=\delta_{i-1}\end{aligned}$$

所以可得 $D_{\delta_i}\to 0$。

定理 1 表明，若迭代终止条件为相邻两次更新生成云的差异足够小，则只需考察云期望向量 Ex_i 和中间向量 G_i 的改变量是否小于给定阈值即可。

需要补充的是，云部分增量式更新的根本目的不是为了增量式求解各端点云，而是云生成的启发式策略中进行不重复随机采样时用于判断迭代的终止条件，因为部分增量式更新具有较快的速度。

4.3.4 云差异的弦度量

相邻两次更新所生成云之间差异的刻画是云

启发式生成策略需解决的又一重要问题。由定理 1 可知，用云期望向量 $\boldsymbol{E}x_i$ 及中间向量 $\boldsymbol{G}_i$ 的改变量来刻画第 i 次迭代后所生成云 C_i 与迭代前的云 C_{i-1} 之差异 D_{C_i} 是合适的。即

$$a_1=\left\|\boldsymbol{E}x_i\right\|_2, a_2=\left\|\boldsymbol{E}x_{i-1}\right\|_2, b_1=\left\|\boldsymbol{G}_i\right\|_2, b_2=\left\|\boldsymbol{G}_{i-1}\right\|_2$$

其中，$\|\bullet\|_2$ 为 l_2 范数，本研究用弦度量定义 D_{C_i} 为

$$D_{C_i}=r((a_1,b_1),(a_2,b_2))=\frac{\left|a_1b_2-b_1a_2\right|}{\sqrt{(a_1^2+b_1^2)\times(a_2^2+b_2^2)}} \tag{10}$$

这种间接度量方式除了具有相邻云之间差异的刻画能力外，其另外 2 个优点在于：①规范性，即 $D_{C_i}\in[0,1]$；②异常值的不敏感性，显然 $\boldsymbol{E}x_i$ 和 $\boldsymbol{G}_i$ 对异常值是敏感的，当异常值出现时，可对弦度量对应的 Riemann 球面做一个适当旋转，此旋转对应着异常值的 l_2 范数的一个变换，变换后的值为非异常值，其优势是保持弦度量不变。限于篇幅，本研究不再深入探讨异常值的检测及处理等细节。

4.3.5　改进的多维逆向云发生器算法

基于大数据随机采样法以及启发式的云生成策略，本文对无确定度的多维逆向正态云发生器[30]进行改进，使其适宜于大数据环境下云的快速生成。改进后的算法如下。

算法 1　大数据多维逆向云发生器 BDMBCG。

输入：子块数目 s，初始抽样率 r_0，云差异阈值 e。

输出：云 $C(\boldsymbol{E}x,\boldsymbol{E}n,\boldsymbol{H}e)$。

1) 初始化：将分块存储在当前云端的数据按 4.3.1 节所述的数据子块划分方式将其分割成 s 个子块，并求每个子块大小 s_0，置 $n=r_0s_0$，置 r 为小于 n 的随机正整数。

2) 进行两次容量分别为 n 和 r 的不重复随机采样，并根据式(6)计算均值向量 $\boldsymbol{E}x_0$ 和 $\boldsymbol{E}x_1$ 以及中间向量 $\boldsymbol{G}_0$ 和 $\boldsymbol{G}_1$，再根据式(10)求云差异 D_C。

3) WHILE　$D_C>e$ 且数据未抽样完时。

4) $\boldsymbol{E}x_0=\boldsymbol{E}x_1$，$\boldsymbol{G}_0=\boldsymbol{G}_1$，$n=n+r$。

5) 执行容量为 r 的不重复随机采样，当所剩样本不足 r 时，抽取剩余样本的 $\frac{1}{2}$。

6) 根据式(7)更新 $\boldsymbol{E}x_1$，并根据式(8)更新 $\boldsymbol{G}_1$。

7) 根据式(10)求云差异 $D_{C'}$。

8) IF　$D_{C'}-D_C>e/2$

9) 产生小于 r 的随机正整数 t，并置 $r=t$；

10) ELSE

11) 产生介于(r, s_0)之间的随机正整数 t，并置 $r=t$；

12) ENE

13) 置 $D_C=D_{C'}$。

14) END　//End While

15) $\boldsymbol{S}^2=\frac{1}{n-1}\boldsymbol{G}_1$，$\boldsymbol{A}=\frac{1}{n}\sum_{k=1}^{n}|\boldsymbol{x}_k-\bar{\boldsymbol{x}}|$。

16) $\boldsymbol{E}x=\boldsymbol{E}x_1$, $\boldsymbol{E}n=\sqrt{\mathrm{p}/2}\boldsymbol{A}$, $\boldsymbol{H}e=\sqrt{\boldsymbol{S}^2-\boldsymbol{E}n\otimes\boldsymbol{E}n}$。

算法 1 的最后两步表明，尽管云部分增量式更新的根本目的不是为了增量式求解各端点云，但却达到了部分增量式求解的目的，因为求解云 $C(\boldsymbol{E}x,\boldsymbol{E}n,\boldsymbol{H}e)$ 时，只重新计算中间量 $\boldsymbol{A}$，其余量直接应用算法在启发式迭代过程中增量更新的值。

注：1)算法 1 在各个云端执行，本研究假设数据 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 作为云端公共变量可直接访问，因此算法输入省略此数据项；2)每个云端数据 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 的容量往往不相等，由于 CCA 要求输入的两组样本容量一致，因此算法执行后还需进行一次随机采样，其操作在小样本容量对应的数据上进行，所抽取样本量为算法 1 执行后获得的两组样本量之差值。

4.4　多维云合并

在式(5)对应的云合并运算中，每次仅能进行一对云加法运算，如果通过反复调用方式每次合并一对云，每合并一次，云的总个数仅减少一个，因为新生成的云还需要加入合并操作，这在云端较多时将增大时间开销，特别在大数据环境下，其效率会遭受质疑；另一方面，式(5)也未顾及 2 个云重要性的差异，在大数据环境中，由于受数据收集或存储策略等差异的影响，不同云端的数据可能存在重要性差异，因此各云端传送到中心云端的云的合并应体现各云端之差异。

针对前述不足，本文借鉴文献[30]用于概念粒度提升的跃升策略的相邻云合并思想，提出了一种适宜于大数据的云合并运算方法。

给定 m 个 p 维云 $C_i(\boldsymbol{E}x_i,\boldsymbol{E}n_i,\boldsymbol{H}e_i)$ $(i=1,2,\mathrm{L},m)$，以及刻画每个云重要度的向量 $\boldsymbol{h}=(h_1,h_2,\mathrm{L},$

$h_m)^{\mathrm{T}}$，$\sum h_i = 1$，记

$$\boldsymbol{M}_x = (\boldsymbol{E}x_1, \boldsymbol{E}x_2, \text{L}, \boldsymbol{E}x_m)$$

$$\boldsymbol{M}_n = (\boldsymbol{E}n_1, \boldsymbol{E}n_2, \text{L}, \boldsymbol{E}n_m)$$

$$\boldsymbol{M}_h = (\boldsymbol{H}e_1, \boldsymbol{H}e_2, \text{L}, \boldsymbol{H}e_m)$$

$\boldsymbol{E}x = (e_x^j)_{p\times 1}, \boldsymbol{E}n = (e_n^j)_{p\times 1}, \boldsymbol{H}e = (h_e^j)_{p\times 1}$。若合并后的云为 $C(\boldsymbol{E}x, \boldsymbol{E}n, \boldsymbol{H}e)$，则有

$$\begin{cases} e_n^j = < \boldsymbol{r}_n', \boldsymbol{h} > \\ e_x^j = < \boldsymbol{r}_x, \boldsymbol{r}_h > / e_n^i \\ h_e^j = < \boldsymbol{r}_h, \boldsymbol{r}_h > / e_n^i \end{cases} \tag{11}$$

其中，$j = 1, 2, \text{L}, p$，$\boldsymbol{r}_n' = \boldsymbol{M}_n'^{\mathrm{T}} \boldsymbol{1}_j$，$\boldsymbol{r}_x = \boldsymbol{M}_x^{\mathrm{T}} \boldsymbol{1}_j$，$\boldsymbol{r}_h = \boldsymbol{M}_h^{\mathrm{T}} \boldsymbol{1}_j$，$\boldsymbol{r}_h = \boldsymbol{h} \otimes \boldsymbol{r}_n'$，而 $\boldsymbol{1}_j$ 为第 j 个元素为 1，其余元素为 0 的 p 维单位列向量。$\boldsymbol{M}_n' = (e_{ni}^{j'})_{p\times m}$ 求解方法为：令 $C_i^j(e_{xi}^j, e_{ni}^j, h_{ei}^j)$ 为第 i 个 p 维云 C_i 的第 j 个维度构成的一维云，其期望曲线方程为 $y_i^j(x)$。设

$$y_i^{j'}(x) = \begin{cases} y_i^j(x), & \text{若 } y_i^j(x) \geq y_k^j(x), k = 1, 2, \text{L}, p \\ 0, & \text{其他} \end{cases}$$

则有

$$e_{ni}^{j'} = \frac{1}{\sqrt{2\text{p}}} \int_U y_i^{j'}(x) \text{d}x$$

其中，U 为第 i 个 p 维云 C_i 的第 j 个维度对应论域，$i = 1, 2, \text{L}, m; j = 1, 2, \text{L}, p$。

与已有方法相比，本文提出的云合并方法呈现出 3 个特点：1)能对各云端传入中心云端的云进行一次性合并；2)云合并中体现了不同云端的重要性差异；3)合并的是多维云，而非一维云。

5 仿真实验及结果分析

5.1 实验数据及仿真云平台

实验涉及 3 个数据集。

1) 带噪声的线性数据集 LN：这是一个合成数据集，数据 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 每个属性来自于线性数据，然后叠加符合高斯分布 $N(1,2)$的样本扰动每个属性值。每次产生的数据包括 10 个维度。

2) 真实数据集 PAMAP2：这是对 18 个不同物理活动进行监视所收集的数据（http://archive.ics.uci.edu/ml/datasets/PAMAP2+Physical+Activity+Monitoring），包括 3 850 505 行记录，含 52 个属性。实验选取的属性为惯性测量单元 IMU（inertial measurement units），前两组实验选取手部 IMU（IMU hand）的前 10 个属性，而第三组实验将手部 IMU 作为一组（包括 17 个属性），而胸部 IMU（IMU chest）作为另一组（包括 17 个属性）。

3) 真实数据集 IDS：网络入侵检测数据集 IDS[31] 记录了网络链接中正常链接和攻击性链接（intrusions or attacks）的行为数据，共包括 494 021 条记录，含 41 个属性。实验选取其中的连续属性（包括 34 个）进行测试，前两组实验选取前 10 个属性；第三组实验将前 12 个属性为一组，其余为另一组。

实验前已删除数据集中具有缺失值的记录，且对每个属性在均值 4 倍方差外的值用均值替换。

CCA 以及多维云发生器对数据约束较少，一般认为，只要总体接近正态分布的实数都可采用。选择 PAMAP2 和 IDS 数据集的理由在于它们是得到大量文献广泛采用的标准数据集，而且其容量较大，已接近仿真实验平台的资源上限。

实验从上述 3 个数据集中选取的每个属性都是总体接近正态分布的实数。图 4 是从 PAMAP2 数据集手部 IMU 中随机挑选出的两列数据(IMU6 和 IMU12）的分布直方图。数据已规范化为均值 0，方差 1。设置了 25 个云端，将数据均分为 25 个相邻块，每个云端分配一块。其中，图 4(a)为总体分布直方图，而图 4(b)、图 4(c)和图 4(d)分别为第 3 号、17 号和 23 号云端中的数据分布直方图。

由图 4 可以看出，不论是总体数据还是分配到各云端的数据都接近正态分布；此外，不同云端的均值偏移不同，且方差范围有所区别，此现象说明 4.4 节研究多维云合并是必要的。笔者在做本实验前还对手部 IMU 其他属性、胸部 IMU 的各属性以及 LN 和 IDS 数据集的连续属性都进行了类似的分布情况观察分析，结果与在 IMU6 和 IMU12 上的观察结果相似，篇幅所限，不再赘述。

因此，尽管所选数据集与真实大数据在容量上有一定的差异，但就仿真而言，数据容量、数据总体分布和各云端的数据分布等都有一定的代表性。

实验在单台微机上通过仿真完成。为仿真数据在各云端的存储，实验为每个云端创建一个文件夹，每个文件夹下存储若干纯文本文件，每个文本文件存储一个数据块。每个实验开始前，先将各数据集切分为相邻块并存储到对应文本文件中。

实验为每个云端启动一个独立线程，所有云端

对应线程并行执行。每个线程从所属云端对应文件夹下读取相应数据，并分配一块内存用于存储相应数据。各线程根据读取的数据生成各端点云。若内存资源不足时，正在读取数据的线程挂起，当内存资源可用时再唤醒。在需计算运行时间的实验中，线程从挂起到唤醒所耗时间忽略。

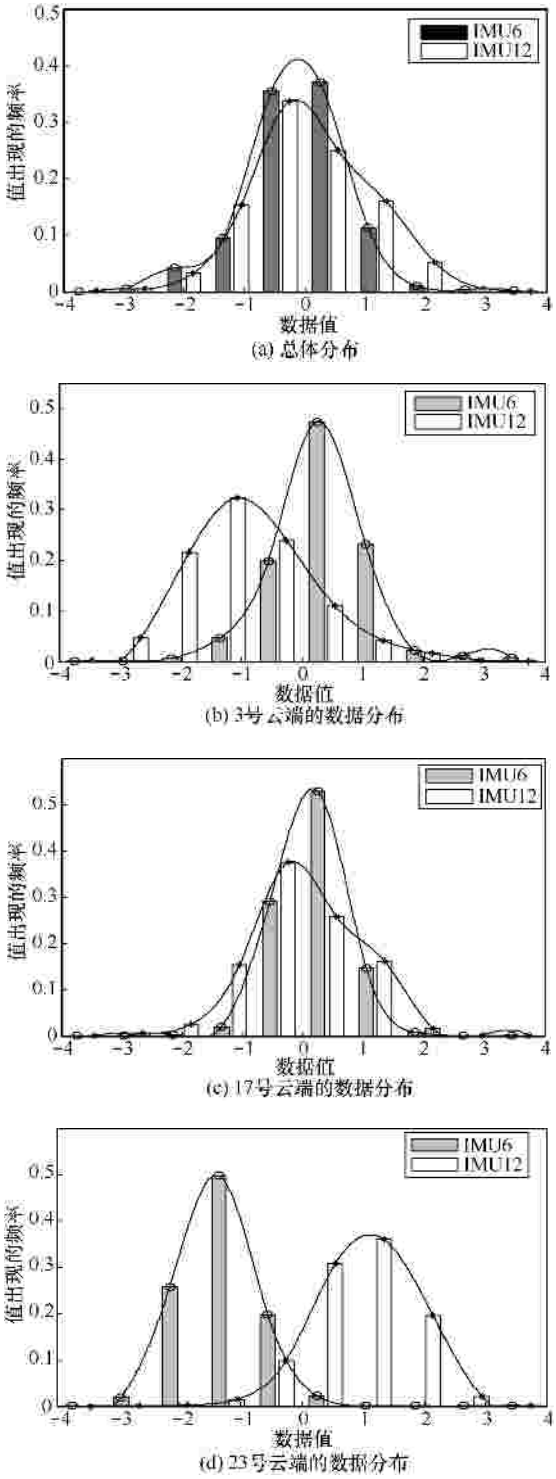


图 4　IMU6 和 IMU12 分布直方图

为中心云端启动一个独立线程，并分配一块互斥访问的内存，用于存储各云端传回的云。当所有端点云都传回后，中心云端对应线程基于此内存块中的云完成云合并、中心云滴产生以及 CCA 运算。

实验通过 C#语言实现，在 Microsoft Visual Studio 2010 Ultimate-CHS 环境中完成，作图工具选用 MATLAB R2011a。实验计算机配置为双核 2.8 GHz CPU、4.0 GB 内存，操作系统为 Windows 7 Professional。

5.2　实验一：各参数对端点云生成的影响

为验证本文改进的多维逆向云发生器 BDMBCG 的有效性，本实验评估各参数对端点云生成的影响。为叙述方便，将改进前的多维逆向云发生器记为 MBCG。由于 BDMBCG 在每个云端运行，因此本组实验设置云端数目为 1，即在 1 个云端观察，并设数据集在每个云端分为 10 块存储。

需考察的参数包括数据子块数目 s、初始抽样率 r_0 和云差异阈值 e。实验将云 $C(\boldsymbol{Ex},\boldsymbol{En},\boldsymbol{He})$ 视为 $R^{p\times 3}$ 上的子空间，p 为维数，用算法改进前计算出的云 $\boldsymbol{C}_1$ 和改进后所得的云 $\boldsymbol{C}_2$ 对应的列子空间 $\boldsymbol{S}_1=\mathrm{col}(\boldsymbol{C}_1)$ 和 $\boldsymbol{S}_2=\mathrm{col}(\boldsymbol{C}_2)$ 的距离 $d(\boldsymbol{S}_1,\boldsymbol{S}_2)$ 作为误差 *error* 的度量，定义为

$$error=d(\boldsymbol{S}_1,\boldsymbol{S}_2)=\left\|\boldsymbol{P}_{\boldsymbol{S}_1}-\boldsymbol{P}_{\boldsymbol{S}_2}\right\|_2 \tag{12}$$

其中，$\boldsymbol{P}_{\boldsymbol{S}_i}=\boldsymbol{C}_i(\boldsymbol{C}_i^{\mathrm{H}}\boldsymbol{C}_i)^{-1}\boldsymbol{C}_i^{\mathrm{H}}$ 为到 $\boldsymbol{C}_i$ 对应的子空间 $\boldsymbol{S}_i$ 上的正交投影算子，$i=1,2$。

需要补充的是，式(12)与式(10)刻画的 2 种云差异的区别：①条件不同，式(12)需求出云期望、熵和超熵后才有意义，而式(10)只需给出云期望向量和中间向量 $\boldsymbol{G}$；②目的不同，式(12)用于直接度量 2 种算法产生的云之间的差异，而式(10)用于间接度量同一算法在云部分增量式更新过程中相邻时刻产生的云之间的差异。由于算法 1 执行后云已经生成，因此用式(12)刻画 BDMBCG 生成的云与 MBCG 生成的云之间的差异是合理的。由上述两点区别得出的结论是，引入式(10)和式(12)是必要的，而且不可用一方代替另一方或交换其位置。

每组实验重复 100 次，以观察不同参数下云的平均差异和计算时间。每次生成 LN 数据 200 000 条记录，每条记录包括 10 维；从 PAMAP2 数据集随机抽取 200 000 条相邻记录，其属性选取为手部 IMU 前 10 个属性；并从 IDS 数据集中随机抽取 200 000 条相邻记录，其属性选取前 10 个连续属性维度。

首先，考察数据子块数目 s 对生成云的影响及计算时间的差异。初始抽样率 $r_0 = 0.35$，云差异阈值 $e = 0.1$。图 5 为误差比较图，而图 6 为 3 个数据集上的平均计算时间比较图。

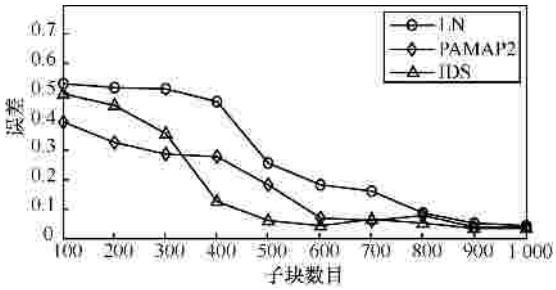


图 5 不同子块数目下所生成云的误差

由图 5 可见，随着子块数目 s 的增大，误差逐渐减小。当 s 增加到 1 000 时，误差已接近 0.05。此现象表明，适当增大子块数目有助于提高计算精度。但图 6 却表明，随着子块数目的增大，BDMBCG 所需时间略有上升。因此在一定精度范围内，子块数目选择适中为宜。此外，真实数据集 PAMAP2 和 IDS 上的误差比合成数据集 LN 上的误差略小。

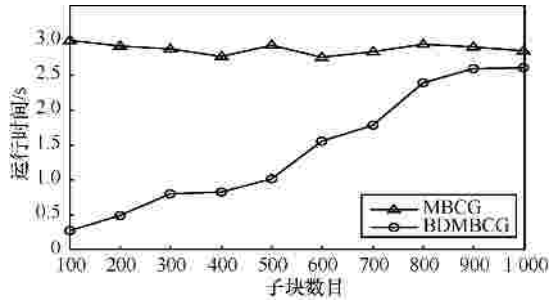


图 6 不同子块数目下平均运行时间

其次，评估初始抽样率 r_0 对生成云的影响。数据子块数目 $s = 400$，云差异阈值 $e = 0.1$。图 7 为误差比较图，而图 8 为不同初始抽样率 r_0 的平均运行时间。

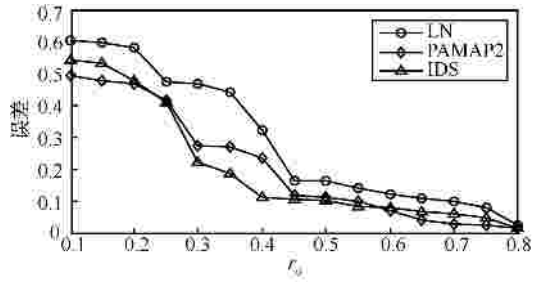


图 7 不同初始抽样率下所生成云的误差

由图 7 可以看出，在 $r_0 \leq 0.3$ 时各数据集上误差都较大；当 r_0 在 0.20~0.45 范围内时，误差下降趋势明显；而此后误差逐渐接近 0.05 左右，且波动较小，其趋势几乎延续到 $r_0 = 0.8$。但是，并不是初始抽样率越大越好，观察图 8 可以发现，当 r_0 变小或增大时，3 个数据集上平均运行时间持续增加。

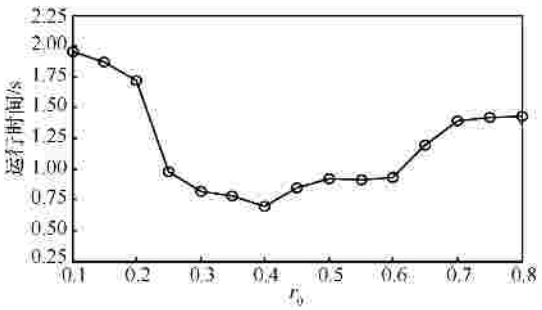


图 8 不同初始抽样率下平均运行时间

再次，观察云差异阈值 e 对生成云的影响及计算时间的差异。初始抽样率 $r_0 = 0.4$，数据子块数目 $s = 400$ 对。图 9 为误差比较图，而图 10 呈现了 3 个数据集上的平均运行时间。由图 9 可以看出，当 $e \geq 0.15$ 时，误差持续增大。图 10 表明，生成云的平均运行时间随着云差异阈值的增大不断减少。结合两图观察发现，当 e 介于[0.08,0.15]时，能获得一个兼顾较低误差和较少运行时间的折中方案。

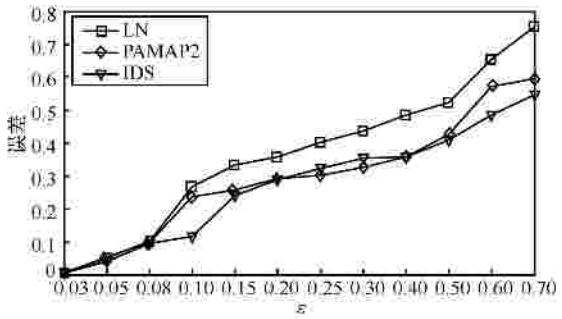


图 9 不同云差异阈值下所生成云的误差

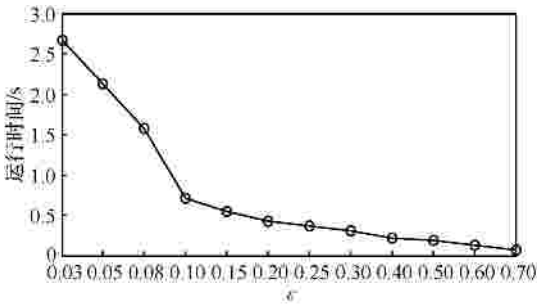


图 10 不同云差异阈值下平均运行时间

5.3　实验二：多维云合并运算的效率分析

本实验将式(5)对应的原始云合并方法（记为“original”）与本文提出的一次性合并多个多维云的云运算方法（如式(11)所示，不妨记为“new”）进行比较，评估不同云端数目对云合并效率的影响。对于式(5)对应的原始云合并，通过反复迭代，每次合并 2 个云，将前一次合并后的云加入当前云的集合再次合并，直至最终合并为一个云为止。

对于同一云端数目 n_{c_i}，实验重复进行 50 次。第 i 次实验中，云重要度皆为 $1/n_{c_i}$。每次实验生成维数为 10 的 LN 数据 $2n_{c_i}\times10^5$ 条记录；并从 PAMAP2 数据集和 IDS 数据集中各随机抽取 2×10^5 条相邻记录 n_{c_i} 次，属性选取与实验一相同。按抽取顺序将数据平均分配到 n_{c_i} 个云端。之后在每个云端并行调用算法 1 的 BDMBCG($s=400$、$r_0=0.3$、$e=0.1$)生成每个端点云，并将生成的云传回中心云端。本实验仅仅评估在中心云端上合成中心云的效率。

图 11 为不同云端数目下，在 3 个数据集上云合并的平均运行时间比较图。由图 11 可以看出，随着云端数目的增大，原始的云合并操作所需时间迅速上升，而本文提出的一次性合并多个多维云的操作所需时间上升幅度却相对较小。此现象表明，本文提出的云合并操作对于所提出的大数据分布式云架构是合适的，云端数目增大并未显著提高云合并的时间开销。

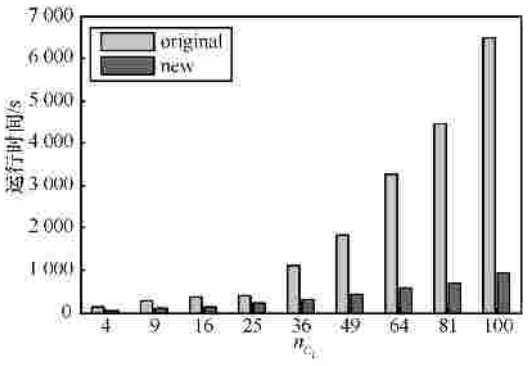


图 11　云合并运行时间比较

5.4　实验三：BDCCA 的有效性评估

为验证本文所提 BDCCA 的有效性，本组实验将 BDCCA 与经典 CCA（记为 NaiveCCA）、ApproxCCA[32]和 LS-CCA[33]进行对比分析，考察不同云滴群大小、不同云端数目以及不同数据总容量下，典型相关系数的精度以及 BDCCA 的执行效率。

典型相关系数的精度用其误差 *error* 刻画。*error* 定义为 NaiveCCA 在原始大数据上所得典型相关系数 r_{NavieCCA} 分别与其他几种方法所得典型相关系数之差的绝对值，即

$$error=\left|r_{\text{NavieCCA}}-r_0\right| \tag{13}$$

其中，r_0 取 r_{BDCCA}、$r_{\text{ApproxCCA}}$ 或 $r_{\text{LS-CCA}}$。r_{BDCCA} 表示 BDCCA 在云滴群上所得的典型相关系数，而 $r_{\text{ApproxCCA}}$ 和 $r_{\text{LS-CCA}}$ 分别表示 ApproxCCA 和 LS-CCA 在原数据上所得的典型相关系数。

基于 BDCCA 求典型相关系数的过程为：对于每个实验，首先在每个云端并行调用算法 1 的 BDMBCG 生成每个端点云，并将生成的云传回中心云端；其次根据式(11)进行云合并；第三采用文献[30]中的多维正向正态云发生器产生中心云滴群 $drop(\boldsymbol{X},\boldsymbol{m}_x)$ 和 $drop(\boldsymbol{Y},\boldsymbol{m}_y)$；最后在 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 上执行 CCA 操作。

本节所有实验在每个云端前两步的参数设置同实验二，且所有实验在数据集 PAMAP2 和 IDS 上进行。在 PAMAP2 数据集上，实验将手部 IMU 作为一组（包括 17 个属性），而胸部 IMU 作为另一组（包括 17 个属性）；而 IDS 数据集则选取前 12 个连续属性为一组，其余连续属性为另一组。

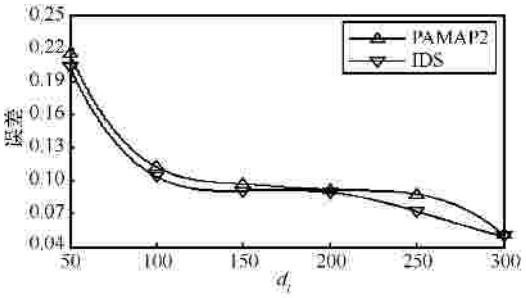
5.4.1　云滴群大小对典型相关系数的影响

本实验设置 25 个云端，每个云端的数据选取方式与实验二相同。对于给定的云滴群大小 d_i，实验重复 30 次，每次都重新挑选数据。对每个典型相关系数，其误差定义如式(13)所示。取各次所得典型相关系数误差的算术平均值作为平均误差。

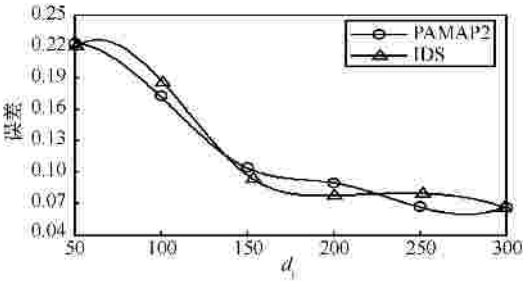
由于 BDCCA 计算典型相关系数是在云滴群上进行的，而其他 CCA 方法则在原始大数据上进行，因此当数据总容量固定后，云滴群的规模并不影响 ApproxCCA 和 LS-CCA 所得典型相关系数的误差，因为在不同云滴群大小下，$r_{\text{ApproxCCA}}$ 和 $r_{\text{LS-CCA}}$ 为常数。故本实验仅仅考察不同云滴群大小下 BDCCA 所得典型相关系数误差的变化情况。

图 12 为不同云滴群大小下前 2 个典型相关系数的平均误差。由图 12 可看出：①随着云滴群规模的增大，前 2 个典型相关系数的误差均逐渐降低，但当云滴群大小超过 150 时，其降低趋势趋于平缓。此现象的启发是，适当增大云滴群规模

有助于降低典型相关系数的误差，但是当其规模增大到一定程度后，再增加云滴数目对于降低误差的贡献并不大。②当云滴群大小超过 100 时，相关系数的误差均较小，其值未超过 0.2，多数在 0.1 范围内。此现象从相关性这一侧面揭示了大数据的稀疏价值特性，即大量数据中蕴含的相关性通过少量云滴即得以刻画，这与本文研究的最初设想是一致的。



(a) 第 1 典型相关系数对应误差



(b) 第 2 典型相关系数对应误差

图 12 不同云滴群大小下典型相关系数误差

5.4.2 云端数目的影响

在数据总容量和云滴群大小均给定的情况下，本实验评估云端数目对典型相关系数的精度及运行时间的影响。与 5.4.1 节的实验相似，当数据总容量固定后，云端规模也不会影响 ApproxCCA 和 LS-CCA 所得典型相关系数的误差，因此本实验关于典型相关系数误差也仅仅考察 BDCCA 所得典型相关系数误差随云端数目变化而变化的情况。误差定义如式(13)所示。

云滴群大小设置为 100。从数据集 PAMAP2 和 IDS 中重复抽取 100 次数据，每次随机抽取 2×10^5 条相邻记录。当云端数目 n_{c_i} 给定后，第 i 个云端分配的记录数目为 $\lfloor 2\times10^7 / n_{c_i} \rfloor$，其中，$\lfloor\bullet\rfloor$ 表示向下取整。实验对不同的云端数目 n_{c_i} 重复 10 次。图 13 为不同云端数目下第 1 典型相关系数的平均误差，而图 14 为不同云端数目下的平均运行时间。

5.4.3 数据容量的影响

本实验考察数据总容量对 BDCCA 所得典型相关系数的精度和运行时间的影响。从数据集 PAMAP2 和 IDS 中重复抽取若干次数据，每次随机抽取 1×10^5 条相邻记录，直至所取数据达到所需容量 n_D 为止。共进行 10 组实验，云端数目设置为 $10^{-6}n_D$，即每个云端分配 1×10^6 条记录。云滴群大小设置为 100。每组实验重复 10 次，取每次所得典型相关系数误差的平均值作为输出误差，而取所有云端的最大运行时间作为 BDCCA 的运行时间。

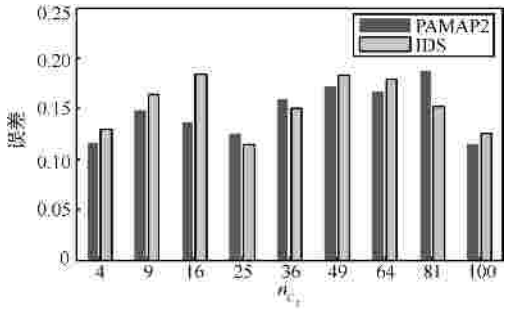


图 13 不同云端数目下典型相关系数的误差

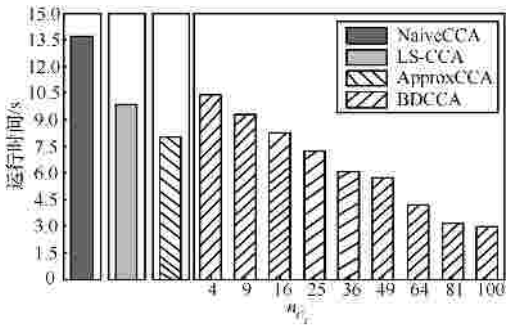


图 14 不同云端数目下的平均运行时间

表 1 为不同总数据容量下前 2 个典型相关系数的平均误差。误差定义如式(13)所示。由表 1 可知：1）从总体上看，BDCCA、ApproxCCA 和 LS-CCA 对应典型相关系数误差都随着数据总容量的增加而上升，但后两者是持续地快速上升，且上升幅度较大，而 BDCCA 在上升过程中存在波动，且上升幅度略小；2）当数据总容量较小时，BDCCA 对应典型相关系数误差略大于 ApproxCCA 和 LS-CCA 对应误差，而当数据总容量较大时，后两者对应误差迅速超过前者对应误差（见表中粗体）。上述现象表明，在数据容量较大的情况下，BDCCA 所得典型相关系数精度相对略高，从这个意义上说，BDCCA 用于大数据分析是适宜的。

表 1　不同数据容量下典型相关系数平均误差

组数	数据总容量（×10^7）											
	第 1 典型系数						第 2 典型系数					
	PAMAP2			IDS			PAMAP2			IDS		
	BDCCA	ApproxCCA	LS-CCA	BDCCA	ApproxCCA	LS-CCA	BDCCA	ApproxCCA	LS-CCA	BDCCA	ApproxCCA	LS-CCA
1	0.098 6	0.067 5	0.031 9	0.125 0	0.077 0	0.020 5	0.064 8	0.030 1	0.009 6	0.125 1	0.086 6	0.083 0
2	0.119 2	**0.136 8**	0.092 5	0.136 9	0.108 3	0.070 8	0.059 7	0.049 7	0.035 0	0.136 4	0.090 8	0.086 6
3	0.119 3	**0.157 4**	0.105 3	0.132 6	**0.155 1**	0.109 7	0.087 2	0.082 6	0.051 4	0.141 9	0.136 2	0.122 5
4	0.121 0	**0.170 9**	**0.153 7**	0.157 5	**0.211 5**	**0.159 0**	0.096 6	**0.109 0**	0.082 0	0.145 3	**0.182 6**	**0.174 5**
5	0.137 1	**0.268 6**	**0.213 0**	0.166 2	**0.315 9**	**0.219 3**	0.102 1	**0.116 7**	0.095 2	0.153 3	**0. 231 8**	**0.194 2**
6	0.127 2	**0.272 9**	**0.225 9**	0.172 3	**0.353 3**	**0.255 0**	0.099 7	**0.213 6**	**0.113 7**	0.162 4	**0.336 3**	**0.279 8**
7	0.129 6	**0.287 0**	**0.228 9**	0.164 7	**0.366 2**	**0.273 3**	0.121 6	**0.231 9**	**0.153 1**	0.159 5	**0.351 9**	**0.301 0**
8	0.133 5	**0.336 0**	**0.270 1**	0.157 5	**0.423 9**	**0.287 1**	0.113 9	**0.297 5**	**0.163 4**	0.161 9	**0.396 0**	**0.311 6**
9	0.124 9	**0.352 1**	**0.284 2**	0.160 5	**0.432 7**	**0.333 2**	0.111 8	**0.319 9**	**0.178 7**	0.163 0	**0.433 9**	**0.329 2**
10	0.145 9	**0.408 7**	**0.314 2**	0.173 9	**0.438 6**	**0.408 1**	0.116 9	**0.365 1**	**0.196 0**	0.179 7	**0.444 5**	**0.342 7**

图 15 为不同容量下的平均运行时间。由图 15 可见，BDCCA 的平均运行时间并未因数据容量的增大而显著增加，但 ApproxCCA、LS-CCA 和 NaiveCCA 的平均运行时间则随着数据容量的增加而呈线性递增趋势。此现象表明，如果数据容量增大时对等地增加云端数目，则 BDCCA 能获得较快的处理速度，这恰是大数据的巨大规模特性所欢迎的。

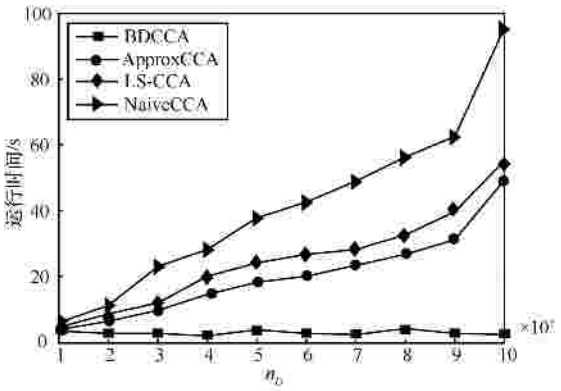


图 15　不同数据总容量下平均运行时间

总之，上述实验结果表明，基于所设计的大数据分布式云架构所提出的 BDCCA 以增加系统资源（即云端）为代价，可获得一定的计算精度和较快的处理速度，这对于大数据快速处理是适宜的。

6　结束语

本文提出了一种面向大数据的 CCA 方法 BDCCA。该方法在容量较小的中心云滴群上进行 CCA 操作，提高了大数据 CCA 的执行效率。为了快速产生中心云滴，首先设计了一种面向大数据挖掘的分布式云架构，为本文大数据存储和计算建立了研究基础；其次重点对多维逆向正态云发生器进行改进，以提高其在大数据环境下产生云的效率；提出了一种一次性合并多个多维云的云合并运算方法，以加快云合并速度。在真实数据集上的实验结果验证了本文方法的合理性和有效性，一方面该方法以增加系统资源（即云端）为代价，可获得一定的计算精度和较快的处理速度；另一方面该方法从相关性这一侧面揭示了大数据的稀疏价值特性。本研究可用于大数据特征融合、机器学习和数据降维等领域。

参考文献：

[1] MINNESOTA M. Big data: science in the petabyte era[J]. Nature, 2008, 455(7209):1-136.

[2] SAKAR C O, KURSUN O. A method for combining mutual information and canonical correlation analysis: predictive mutual information and its use in feature selection[J]. Expert Systems wi h Applications, 2012, 39(3):3333-3344.

[3] OLCAY K, ETHEM A, OLEG V, *et al*. Canonical correlation analysis using within-class coupling[J]. Pattern Recognition Letters, 2011, 32(2): 134-144.

[4] KAMALIKA C, SHAM M K, KAREN L, *et al*. Multi-view clustering via canonical correlation analysis[A]. Proc of the 26th International Conference on Machine Learning[C]. New York, ACM, USA, 2009. 129-136.

[5] **杨静, 李文平, 张健沛. 基于秩 2 更新的多维数据流典型相关跟踪算法**[J]. **电子学报**, 2012, 40(9):1765-1774.
YANG J, LI W P, ZHANG J P. A tracking algorithm based on rank two modifications for canonical correlation analysis of mu idimensional

data streams[J]. Acta Electronica Sinica, 2012, 40(9):1765-1774.

[6] **顾鑫, 徐正全, 刘进. 基于云理论的可信研究及展望**[J]. **通信学报**, 2011, 32(7):176-181.
GU X, XU Z Q, LIU J. Review of cloud based trust model[J]. Journal on Communications, 2011, 32(7):176-181.

[7] **黄海生, 王汝传. 基于隶属云理论的主观信任评估模型研究**[J]. **通信学报**, 2008,29(4):13-19.
HUANG H S, WANG R C. Subjective trust evaluation model based on membership cloud theory[J]. Journal on Communications, 2008, 29(4): 13-19.

[8] **蒋嵘, 李德毅. 基于形态表示的时间序列相似性搜索**[J]. **计算机研究与发展**, 2000, 37(5):601-608.
JIANG R, LI D Y. Similarity search based on shape representation in time-series data sets[J]. Journal of Computer Research & Development, 2000, 37(5):601-608.

[9] **许凯, 秦昆, 黄伯和等. 基于云模型的图像区域分割方法**[J]. **中国图象图形学报**, 2010, 15(5):757-763.
XU K, QIN K, HUANG B H, *et al.* A new method of region based on image segmentation based on cloud model[J]. Journal of Image and Grphics, 2010, 15(5):757-763.

[10] HOTELLING H. Relations between two sets of variates[J]. Biometrika, 1936, 28(3):321-377.

[11] **彭岩, 张道强. 半监督典型相关分析算法**[J]. **软件学报**, 2008, 19(11):2822-2832.
PENG Y, ZHANG D Q. Semi-supervised canonical correlation analysis algorithm[J]. Journal of Software, 2008, 19(11):2822-2832.

[12] **顾晶晶, 陈松灿, 庄毅. 用局部保持典型相关分析定位无线传感器网络节点**[J]. **软件学报**, 2010, 21(11):2883-2891.
GU J J, CHEN S C, ZHUANG Y. Localization in wireless sensor network using locality preserving canonical correlation analysis[J]. Journal of Software, 2010, 21(11):2883-2891.

[13] LI D Y, HAN J W. Knowledge representation and discovery based on linguistic atoms[J]. Knowledge-based Systems, 1998, 7(10):431-440.

[14] PHILIP R. Big Data Analytics[R]. TDWI Best Parctices Report, 2011. 1-38.

[15] BENJAMIN H B, MARK R B, KEITH A S, *et al.* Large-scale electrophysiology: acquisition, comprression, encryption, storage of big data[J]. Journal of Neurosience Methods, 2009, 180(1):185-192.

[16] ARONOVA E, BAKER K, ORESKES N. Big science and big dat in biology: from the international geophysical year through the international biological program to the long term ecological research (LTER) network[J]. Historical Studies in the Natural Sciences, 2010, 40(8): 183-224.

[17] WERNER C. Scienti?c perspectivism: a philosopher of science's response to the challenge of big data biology[J]. Studies in History and Philosophy of Biological and Biomedical Sciences, 2012, 43(1):69-80.

[18] ALFREDO C, YEOL S, KAREN C D. Analytics over largescale multidimensional data: the big data revolution[A]. Proc of the DOLAP'11 [C]. Glasgow, 2011. 101-103.

[19] STEVEN C H H, WANG J L, ZHAO P L, *et al.* Online feature selection for mining big data[A]. Proc of the Big-Mine'12[C]. New York: ACM, USA, 2012. 93-100.

[20] SIMON B, DUODUO L. On clusterization of "big data" streams[A]. Proc of the 3rd International Conference on Computing Geospatial Research and Applications[C]. New York:ACM, USA, 2012.1-6.

[21] JOHN L. Parallel machine learning on big data[J]. XRDS, 2012, 19(1): 60-62.

[22] THOMAS C, PEGGY H, MELANIE M, *et al.* Building a big data research program at a small university[J]. JCSC, 2012, 28(2):95-102.

[23] YU C, CHENG J Q, FLORIN R. GLADE: big data analytics made easy[A]. Proc of the SIGMOD'12[C]. New York: ACM, USA, 2012. 697-700.

[24] KYUSEOK S. MapReduce algorithms for big data analysis[A]. Proc of the 38th International Conference on Very Large Data Bases (VLDB)[C]. New York: ACM, USA, 2012. 2016-2017.

[25] JENS D, JORGE A. Efficient big data processing in hadoop MapReduce[A]. Proc of the 38th International Conference on Very Large Data Bases(VLDB)[C]. New York: USA, ACM, 2012. 2014-2015.

[26] DIVYAKANT A, SUDIPTO D, AMR E A. Big data and cloud computing: current state and future opportunities[A]. Proc of the EDBT 2011[C]. New York: ACM, USA, 2011. 530-533.

[27] XU H Q, LI Z, GUO S M, *et al.* CloudVista: interactive and economical visual cluster analysis for big data in the cloud[A]. Proc of the 38th International Conference on very Large Data Bases(VLDB)[C]. New York: USA, ACM, 2012. 1886-1889.

[28] COLIN T, DIGITAL P. Big data security[J]. Network Security, 2012, 7(2):5-8.

[29] SOTIRIS K. Combining bagging, boosting, rotation forest and random subspace methods[J]. Artificial Intelligence Review, 2011, 35(3):223-240.

[30] **李德毅, 杜鹢. 不确定性人工智能**[M]. **北京: 国防工业出版社**, 2005. 224-227.
LI D Y, DU Y. Artificial Intelligence with Uncertainty[M]. Beijing: National Defence Industry Press, 2005. 224-227.

[31] TAVALLAEE M, BAGHERI E, LU W, *et al.* A detailed analysis of the KDD CUP 99 data set[A]. Proc of the Second IEEE International Conference on Computational Intelligence for Security and Defense Applications[C]. Ottawa, Canada, 2009. 53-58.

[32] WANG Y L, ZHANG G X, QIAN J B. ApproxCCA: an approximate correlation analysis algorithm for multidimensional data streams[J]. Knowledge-Based Systems, 2011, 24(7):952-962.

[33] SUN L, JI S W. Canonical correlation analysis for multilabel classification: a least-squares formulation, extensions, and analysis[J]. IEEE Transactions on Pattern Analysis and Machine Intellige 2011, 33(1):194-200.

作者简介：

杨静（1962-），**女，黑龙江哈尔滨人，哈尔滨工程大学教授、博士生导师，主要研究方向为数据库理论与应用、数据挖掘技术、知识库系统、软件理论等。**

李文平[通信作者]（1979-），**男，贵州大方人，哈尔滨工程大学博士生，主要研究方向为数据流、隐私保护、自然计算。**E-mail:liwenping@hrbeu.edu.cn。

张健沛（1956-），**男，黑龙江哈尔滨人，哈尔滨工程大学教授、博士生导师，主要研究方向为数据库理论与应用、数据挖掘、软件理论等。**