

文章编号:1001-9081(2012)03-0629-05

doi:10.3724/SP.J.1087.2012.00629

面向高速数据流的集成分类器算法

李 南^{1,2}, 郭躬德^{1,2*}

(1. 福建师范大学 数学与计算机科学学院, 福州 350007;
2. 福建师范大学 网络安全与密码技术重点实验室, 福州 350007)
(*通信作者电子邮箱 ggd@fjnu.edu.cn)

摘要: 数据流挖掘要求算法在占用少量内存空间的前提下快速地处理数据并且自适应概念漂移, 据此提出一种面向高速数据流的集成分类器算法。该算法将原始数据流沿着时间轴划分为若干数据块后, 在各个数据块上计算所有类别的中心点和对应的子空间; 此后将各个数据块上每个类别的中心点和对应的子空间集成作为分类模型, 并利用统计理论的相关知识检测概念漂移, 动态地调整模型。实验结果表明, 该方法能够在自适应数据流概念漂移的前提下对数据流进行快速的分类, 并得到较好的分类效果。

关键词: 概念漂移; 数据流; 子空间; 分类; 集成

中图分类号: TP18; TP311 **文献标志码:**A

Ensemble classification algorithm for high speed data stream

LI Nan^{1,2}, GUO Gong-de^{1,2*}

(1. School of Mathematics and Computer Science, Fujian Normal University, Fuzhou Fujian 350007, China;
2. Key Laboratory of Network Security and Cryptography, Fujian Normal University, Fuzhou Fujian 350007, China)

Abstract: The algorithms for mining data streams have to make fast response and adapt to the concept drift at the premise of light demands on memory resources. This paper proposed an ensemble classification algorithm for high speed data stream. After dividing a given data stream into several data blocks, it computed the central point and subspace for every class on each block which were integrated as the classification model. Meanwhile, it made use of statistics to detect concept drift. The experimental results show that the proposed method not only classifies the data stream fast and adapt to the concept drift with higher speed, but also has a better classification performance.

Key words: concept drift; data stream; subspace; classification; integration

0 引言

随着信息产业的发展, 超市交易、电信等众多应用领域每天都产生大量的数据流, 其中蕴含着丰富的有价值的知识有待挖掘, 近年来已成为数据挖掘领域的一个研究热点。由于数据流具有快速性、无限性和实时性的特点^[1], 使得传统的挖掘算法显得有些力不从心。同时, 数据流中隐含的概念或知识可能会随着时间的推移或环境的改变而发生变化, 即 1996 年 Widmer 和 Kubat^[2]提出的概念漂移问题。因此, 数据流挖掘要求算法能在有限的计算时间和内存资源内完成挖掘任务, 并且根据当前的概念自适应地改变模型^[3]。

目前, 处理数据流上概念漂移的方法有 3 种^[4]: 实例选择、实例加权和集成学习。Hansen 等^[5]证明使用集成分类器方法比仅使用单个分类器方法具有更好的适应性和精确性。Wang 等^[6]提出了一个集成学习的通用框架用于挖掘概念漂移数据流。Street 等^[7]提出一个可以自适应数据流概念漂移的集成分类器算法 (Streaming Ensemble Algorithm, SEA), 展示了集成学习的有效性。此后, 许多学者深入研究了集成分类器的权值设计^[8–10]以及集成策略^[11–13]。

然而, 上述已存在的数据流分类模型不仅构建模型耗时多, 而且面临着同一个问题: 当数据流中只有少部分类别发生

概念漂移时, 仍必须抛弃现有的整个集成分类模型进行重建以适应新的概念, 降低了分类效率。针对以上问题, 本文提出了一种新颖的面向高速数据流的集成分类算法 (简称 ECA)。

1 相关工作

1.1 ECA 基分类器构建

最近邻分类是一种已经被众多学者广泛研究的有监督的机器学习方法。经典的 k-最近邻 (k-Nearest Neighbor, KNN) 算法^[14]由于简单但颇为有效被列为十大数据挖掘算法之一^[15]。然而, 其面临分类速度较慢和 k 值难以确定的问题。为了解决分类效率低的问题, 近期有学者通过将同类别数据聚类生成若干关键数据以减少要搜索的近邻数, 在不失分类精度的前提下提高了分类速度^[16]。受其启发, 本文提出一种基于子空间中心点的分类算法作为 ECA 的基分类器。

设训练数据集 X 由 N 个样本组成, 即 $X = \{(x_1, y_1), (x_2, y_2), \dots, (x_i, y_i), \dots, (x_n, y_n)\}$ 。其中 x_i 表示由 D 个属性构成的第 i 个样本, 即 $x_i = (x_{i1}, x_{i2}, \dots, x_{iD})$; $y_i \in \{1, 2, \dots, K\}$ 表示 x_i 的类别, $K (K > 1)$ 表示训练集中包含的类别个数。为了进一步减少要搜索的近邻数目以提高分类效率, 将所有同类别数据组成一个中心点, 记为: $\text{Center}_k = \{c_{k1}, c_{k2}, \dots, c_{kD}\}$ 。

收稿日期:2011-08-22;修回日期:2011-11-20。

基金项目:国家自然科学基金资助项目(61070062, 61175123); 福建高校产学合作科技重大项目(2010H6007)。

作者简介:李南(1987-), 男, 福建福州人, 硕士研究生, 主要研究方向:信息融合、数据流挖掘; 郭躬德(1965-), 男, 福建龙岩人, 教授, 博士, 主要研究方向:数据挖掘、机器学习。

$$\text{Center}_k = \frac{1}{\text{Num}(k)} \sum_{y_i=k} x_i \quad (1)$$

其中 $\text{Num}(k)$ 表示第 k 类的样本数目。

显然,将所有同类别数据仅用一个中心数据来表示不仅提升了模型对噪声的鲁棒性,同时大大提高了分类效率。但值得注意的是,将所有同类别数据仅用一个中心数据来表示分类时容易受到数据离散程度的影响,图 1 就是个例子。在二维空间上空心的椭圆和矩形分别代表两类不同的样本,其中心点各自用实心的椭圆和矩形表示。如果简单地利用各自的中心点来代表所有同类数据,根据测试样本在全空间上距离两类中心点的距离来进行分类(相当于用图中的虚线作为分类标准),显然不能正确地代表数据的分布情况,分类处在全空间类别边界上的点时精度受到影响。因此,有必要对其进行进一步改进。

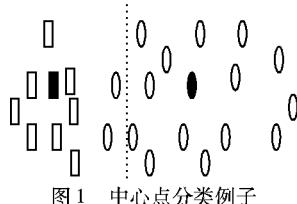


图 1 中心点分类例子

数据空间中往往存在许多不相关的属性,在全空间上表现为同类别的是“离散的”,只在某些低维的子空间上是“密集的”^[17]。为了减少数据离散程度对分类模型所造成的影响,我们将测试样本投影到每个类别所在的子空间上,即利用加权的欧氏距离来衡量测试样本与各个类别中心点的距离。算法基于软子空间聚类^[18]的普遍假设:“维度权重大小与同类数据点投影到该维度上的分布离散程度成反比”的思想来建立每个类别所在的子空间,记为:

$$\begin{aligned} \text{Weight}_k &= \begin{bmatrix} w_{k1} & 0 & 0 & \cdots & 0 \\ 0 & w_{k2} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & w_{kd} \end{bmatrix} \\ w_{kj} &= \left(\sum_{m=1}^d \left(\frac{\sum_{y_i=k} [(x_{ij} - c_{kj})^2 + \delta]}{\sum_{y_i=k} [(x_{im} - c_{km})^2 + \delta]} \right)^2 \right)^{-1} \end{aligned} \quad (2)$$

其中: δ 是为避免分母为 0 而引入的一个很小的数值,根据软子空间聚类算法 FWKM^[19],本文取 $\delta = 10^{-4}$ 。

综上,ECA 基分类器构建算法如下:

算法 BaseT。

输入 训练集 $TrainInstances$ 。

输出 $TrainInstances$ 中每个类别的中心点和权重(即一个基分类器)。

- 1) for $TrainInstances$ 中的每一个类别 k
- 2) 利用式(1)计算其中心点 Center_k
- 3) 利用式(2)计算其子空间 Weight_k
- 4) end for

算法 BaseC。

输入 训练好的一个基分类器,测试样本 x_i 。

输出 x_i 的类别 y_i 。

- 1) for 基分类器中的每一个类别 k
- 2) 将 x_i 投影到 k 类所在的投影子空间上,计算其与中心点

$$\text{dist}(x_i, \text{Center}_k) = \sqrt{\sum_{d=1}^D w_{kd} (x_{id} - \text{Center}_{kd})^2}$$

- 3) 输出距离 x_i 最近的中心点的类别号
- 4) end for

1.2 基分类器算法空间复杂度分析

设在大小为 S 、类别个数为 K 的数据块上构建基分类器,根据上述算法流程,其所需的存储空间为 $O(K)$,即存储 K 个中心点数据,通常 $K \ll S$ 并且 K 与 S 无关。同时,求取中心和权重的过程时间复杂度与 S 的大小成线性关系,即算法时间复杂度为 $O(S)$ 。综上,相对于数据块大小 S ,该基分类器算法具有常数的空间复杂度和线性的时间复杂度。因此,该算法适合作为数据流集成分类器的基分类器算法。

2 ECA 的设计与分析

本章先介绍 ECA 分类模型的概念漂移检测机制,然后对算法进行具体描述。算法使用滑动窗口模型,将数据流沿时间轴组织成固定大小 S 的数据块序列,每个数据块用 D_1, D_2, \dots, D_n 表示。

2.1 漂移检测

本文采用假设检验中 χ^2 拟合检验的原理来进行漂移检测。其基本思想是如果相邻两个数据块内关于同一类别的数据的分类精度在一定的显著性水平下有显著改变,那么就有理由认为新数据块上的此类数据概念发生变化,需要重构该类别的分类模型。

χ^2 拟合检验的原理^[20]是:当总体的分布未知时,根据样本 X_1, X_2, \dots, X_z 来检验关于总体分布的假设“ H_0 : 总体 X 的分布函数是 $F(x)$ ”。设 X 的取值范围为 A_1, A_2, \dots, A_z 。以 $f_i (i = 1, 2, \dots, z)$ 记录 n 个样本观测值 x_1, x_2, \dots, x_n 落在 A_i 的个数, $p_i (i = 1, 2, \dots, z)$ 为根据 H_0 所假设的 X 的分布函数来计算事件 A_i 的概率。那么若样本个数 n 充分大,则当 H_0 为真时,统计量 $\chi^2 = \sum_{i=1}^z \frac{f_i^2}{np_i} - n$ 近似地服从 $\chi^2(z-1)$ 分布。即当样本的观测值使 χ^2 值有 $\chi^2 \geq \chi^2_{\alpha}(z-1)$,则在显著性水平 α 下拒绝 H_0 ;否则就接受 H_0 。

ECA 使用上述的原理检测概念漂移,依次对当前数据块上每个类别的分类情况进行假设检验,设置 $z = 2$ (分类正确与否两种情况), H_0 为前若干个概念平稳的数据块上该类别的平均分类情况,显著性水平 α 取 0.05。若在新数据块上 $\chi^2 \geq \chi^2_{\alpha}(z-1)$,则该类别分类精度发生显著性改变,从而说明发生了概念漂移;反之认为概念分布平稳。

2.2 ECA 描述

ECA 由为每个类别保存的在不同数据块上建立的多个中心点和对应的子空间组成集成模型。同时,当使用数理统计的相关知识检测到数据流的少部分类别发生概念漂移时,无需像现有的集成分类算法^[8-13]一样,耗时耗力地重构整个集成分类模型,降低算法的分类效率。新算法只需将新数据块上建立的符合新概念的该类别的中心点和对应的子空间替换原有分类模型中的即可,符合数据流要求算法能快速处理数据并且自适应概念漂移的特点。

算法流程具体如下:1) 当新的数据块到来时,ECA 先利用现有的分类模型,计算新数据块中各待分类样本与每个类别的距离(距离采用在相应的子空间上待分类样本与为每一个类别保留的不超过 Num 个在各个数据块上建立的中心点的平均距离作为待分类样本距离此类别的距离),选取距离

最近的类别作为待分类样本的类别。2)在新数据块上利用BaseT算法建立该数据块上各类样本的中心点和对应的子空间。3)根据分类情况检查各类别是否发生概念漂移。4)一旦检测出某个类别发生概念漂移,那么删除原有分类模型中所有为该类别保存的中心点和对应的子空间,保存新数据块上建立的该类别的中心点和对应的子空间。如果没有发生概念漂移,先保存新数据块上建立的该类别的中心点和相应的子空间,再判断原有分类模型中为该类别保存的中心点和相应的子空间个数是否超过Num,如果超过,则删除最早建立的那个数据块上构建的中心点和对应的子空间。算法每个数据块上对应的处理流程如下:

算法 ECA。

输入 集成分类器 EC_{n-1} ,当前数据块 D_n ,为各类别保存的中心点和相应的子空间容量Num。

输出 当前分类模型 EC_n 。

- 1) if EC_{n-1} 为空
- 2) 使用 BaseT 算法,在 D_n 上建立各类别的中心点和对应的子空间并将其加入 EC_{n-1} 中组成 EC_n , return
- 3) end if
- 4) for D_n 中的每一个样本
- 5) 计算其与每个类别的距离(采用在相应的子空间上该样本与 EC_{n-1} 中为每一个类别保留的在各个数据块上建立的中心点的平均距离作为待分类样本距离此类别的距离),选取距离最近的类别作为其类别
- 6) end for
- 7) 使用 BaseT 算法,在 D_n 上建立各类别的中心点和对应的子空间
- 8) for D_n 中的每一个类别 k
- 9) 采用 2.1 节中介绍的概念漂移检测机制判断该类别是否发生概念漂移
- 10) if 某类别发生概念漂移
- 11) 删除 EC_{n-1} 中保存的发生概念漂移类别的中心点和对应的子空间
- 12) 存储在 D_n 上建立的该类别的中心点和对应的子空间
- 13) end if
- 14) if 没有发生概念漂移 && EC_{n-1} 中为该类别保存的中心点和对应的子空间数目少于 Num
- 15) 直接保存 D_n 上建立的该类别的中心点和对应的子空间
- 16) end if
- 17) if 没有发生概念漂移 && EC_{n-1} 中为该类别保存的中心点和对应的子空间数目不少于 Num
- 18) 用 D_n 上建立的该类别的中心点和对应的子空间替换 EC_{n-1} 中建立时间最早的该类别的中心点和对应的子空间
- 19) end if
- 20) end for

其中: EC 表示 ECA 集成分类模型, EC_n 表示第 n 个数据块时的集成分类模型。

通过算法流程可看出:ECA 中利用现有分类模型对当前数据块中的数据进行分类,显然只需要相对于数据块大小线性的时间复杂度,其余时间耗费在使用 BaseT 算法在新数据块上建立新的分类模型以对现有模型进行更新。在大小为 S 、类别个数为 K 的数据块上,BaseT 算法为每类计算其中心点及其对应的子空间需要相对于该类别样本数目线性的时间复杂度,因此整体的时间复杂度为 $O(KS)$,通常 $K \ll S$ 并且 K 是独立于 S 的常数。因此,相对于数据块大小 S ,利用 BaseT 算法

对分类模型进行更新具有线性的时间复杂度。综上所述,ECA 具有相对于数据块大小 S 线性的时间复杂度,适合数据流分类模型快速处理的要求。

3 实验分析与讨论

为了评估 ECA 的性能,我们在分别在真实数据集和实验数据集上对算法的精确度和分类效率进行实验。实验环境如下:2.6 GHz CPU 和 2 GB RAM;操作系统为 Windows XP;开发环境为基于 Java 语言的 Weka 平台,编译运行环境为 JDK 1.5。

3.1 使用的算法

为了验证本文算法的有效性,对比算法使用经典的 SEA、目前比较流行的实例加权集成分类器(Example-Weight Algorithm for Mining Data Streams, EWAMDS)算法^[8]以及分类器动态集成的 DWM(Dynamic Weighted Majority)算法^[11]。实验中各种算法的具体参数设置分别参照文献[7~8,11]中的实验参数,ECA 中为各类别保存最多不超过 5 个在各数据块上建立的中心点和对应的子空间个数,即 $Num = 5$ 。

3.2 数据集

分别在以下两个数据集上进行实验以检验 ECA 解决数据流分类问题的有效性。

1) 移动超平面(Hyperplane)^[21]:一个 d 维超平面上的样本满足形式 $\sum_{i=1}^d a_i x_i = a_0$ 。在实验中,取 d 为 100,并且随机产生 3 个不同的权重集合。实验使用 30 000 条数据,蕴含 3 个概念,2 次漂移。其中每个概念含有 10 000 条样本,并包含 5% 的噪声样本。

2) 20-Newsgroups (<http://mlg.ucd.ie/files/dataset>):一个常用的文本数据集,它是由 20 个不同新闻组的文档组成。本文使用的数据集是 20-Newsgroups 来自同一个新闻组的部分样本集合,一共分为 6 类:med、baseball、autos、motor、space 和 politics。实验中随机抽取了 4 498 条样本,各类分布情况见表 1 所示,每个样本包含 500 个特征属性。为了消除文档的长度差异带来的影响,数据事先进行了单位向量长度变换。为了模拟一个多类别漂移的情况,以验证算法对真实复杂数据中出现新类问题的快速适应性以及对多类分类问题的处理能力,将数据集划分为两大块:在第一大块数据中,只有 med、baseball、autos 和 motor 4 类;在第二大块数据中,淘汰了 motor 类的数据,并添加了 space 和 politics 两个新的类别。

表 1 20-Newsgroups 中各类分布

类别	实例个数	类别	实例个数
med	1 162	motor	600
baseball	1 162	space	562
autos	450	politics	562

3.3 实验结果与分析

3.3.1 移动超平面数据集

各种算法在移动超平面数据集上每个数据块上的精度对比结果如图 2 所示。由于移动超平面样本个数较多,将数据块大小设置为 500。

从图 2 可看出:在第 20 和 40 个数据块时,由于数据出现概念漂移的情况,各种算法的分类精度骤然下降。但是,随着

漂移数据的增多,分类器逐渐适应了新的概念,分类精度也恢复到了原先的水平。由于 SEA 使用 C4.5 作为基分类器,在处理维度较高的数值型属性数据时分类精度会受到影响,因而分类效果最差。同时,我们可以看出,在大部分情况下,ECA 分类精度优于 DWM 算法,和 EWAMDS 算法相当。此外,各种算法在移动超平面数据集上的处理时间如图 3 所示。从图 3 可看出:由于基分类器构造方式简单,ECA 上在处理时间上对比其他 3 种算法具有相当的优势。

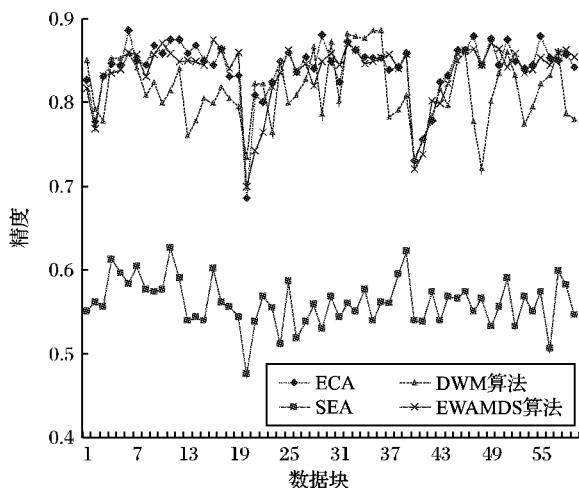


图 2 含 5% 噪声的移动超平面数据流上的分类精度比较

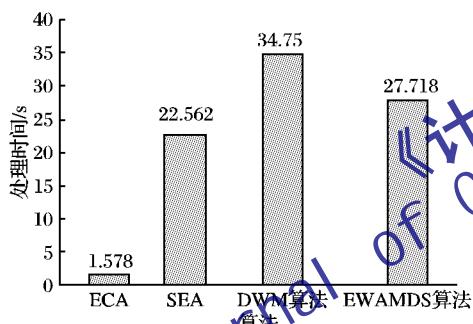


图 3 含 5% 噪声的移动超平面数据流上的处理时间比较

3.3.2 20-Newsgroups 数据集

由于移动超平面是一个二分类问题,为了验证 ECA 在真实复杂结构数据流中面对出现新类问题的快速适应性以及对多类分类问题的处理能力,在 20-Newsgroups 数据集中上进行了测试,此次测试将数据块大小设置为 250。各种算法在每个数据块上的精度对比如图 4 所示。从图 4 可看出:在第 8 个数据块时,由于新类别数据的出现旧类别数据的消失,原有的分类模型已经不适应新数据块的概念,因而各种算法的分类精度出现不同程度的下降。由于 DWM 算法仅根据分类模型中各基分类器的累积错误动态地删除和新建基分类器,因而分类精度降低的幅度最小。同时,在概念稳定以后,ECA 的分类精度高于其他 3 种算法。此外,各种算法对 20-Newsgroups 数据流中每类数据的分类正确率见表 2 所示。从表 2 可看出:对于 6 种类别的判断,ECA 都具有较高的分类精度。

各种算法在 20-Newsgroups 数据流上的处理时间如图 5 所示。虽然只有部分类别发生概念漂移,3 种对比算法仍需重建整个分类模型,无需重建整个模型、基分类器构建简单的 ECA 在处理时间上具有明显的优势。同时 EWAMDS 算法需要为新数据块上的每个样本计算其相应的权重,故需要最长

的处理时间。

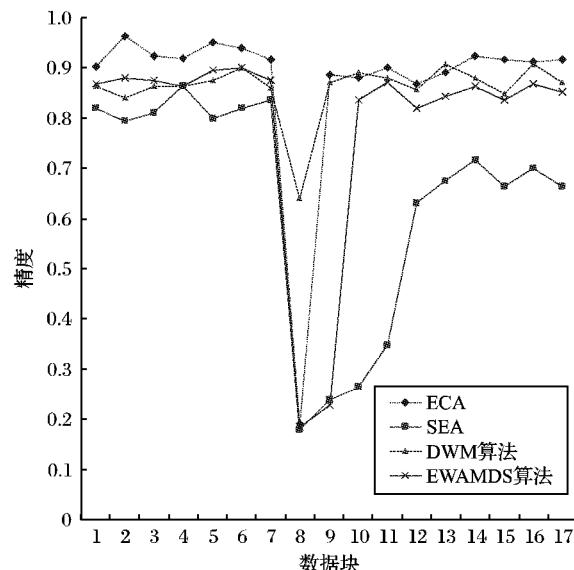


图 4 20-Newsgroups 数据流上的分类精度比较

表 2 在 20-Newsgroups 数据流上每类的分类正确率 %

类别	ECA	SEA	DWM 算法	EWAMDS 算法
med	89.7	81.1	85.5	81.9
baseball	94.8	71.1	91.4	89.3
auto	68.0	53.1	76.7	52.9
motor	96.5	81.7	87.6	91.7
space	82.6	42.5	86.7	70.1
politics	81.3	38.6	87.2	73.8
平均	87.8	65.7	86.8	79.8

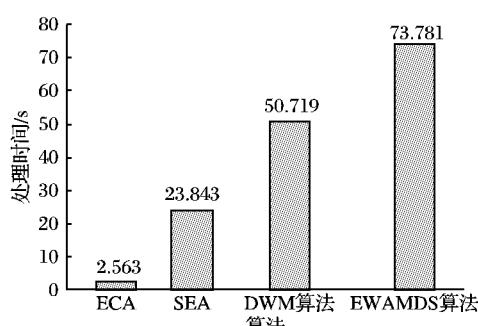


图 5 20-Newsgroups 数据流上的处理时间比较

4 结语

本文针对现有集成分类方法构建分类模型耗时多,在数据流仅部分类别发生概念漂移时仍需重建整个分类模型,分类效率低的缺点,提出一种新的线性时间复杂度的集成分类算法。新算法在部分类别发生概念漂移时仅需重建相应部分的分类模型,从而提高分类效率。在移动超平面和 20-Newsgroups 数据流上的实验表明,与经典的 SEA、当前比较流行的 EWAMDS 算法和 DWM 算法相比,新算法能够在自适应概念漂移的情况下对数据流进行快速分类,并得到较好的分类效果。下一步工作的重点是研究基分类器的建立方法,从而进一步提高分类性能。

参考文献:

- [1] 李燕,张玉红,胡学钢. 基于 C4.5 和 NB 混合模型的数据流分类算法[J]. 计算机科学,2010,37(12):138-142.
- [2] WIDMER G, KUBAT M. Learning in the presence of concept drift

- and hidden contexts [J]. Machine Learning, 1996, 23(1): 69 – 101.
- [3] 王黎明, 周驰. 自适应概念漂移的在线集成分类器[J]. 计算机工程, 2011, 37(5): 74 – 76.
- [4] TSYMBAL A, PECHENIZKIY M, CUNNINGHAM P, et al. Dynamic integration of classifiers for handling concept drift [J]. Information Fusion, 2008, 9(1): 56 – 68.
- [5] HANSEN L K, SALAMON P. Neural network ensemble [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1990, 12(10): 993 – 1001.
- [6] WANG H, FAN W, YU P, et al. Mining concept drifting data streams using ensemble classifiers [C]// KDD'03: Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2003: 226 – 235.
- [7] STREET W, KIM Y. A Streaming Ensemble Algorithm (SEA) for large-scale classification [C]// KDD'01: Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2001: 77 – 382.
- [8] 胡学刚, 潘春香. 基于实例加权方法的概念漂移问题研究[J]. 计算机工程与应用, 2008, 44(21): 188 – 190.
- [9] 欧阳震铮, 罗建书, 胡东敏, 等. 一种不平衡数据流集成分类模型[J]. 电子学报, 2010, 38(1): 184 – 189.
- [10] 张健沛, 杨显飞, 杨静. 面向高速数据流的偏倚抽样集合分类器[J]. 北京邮电大学学报, 2010, 33(4): 44 – 48.
- [11] JEREMY Z K, MARCUS A M. Dynamic weighted majority: An ensemble method for drifting concepts [J]. Journal of Machine Research, 2007, 8(12): 2755 – 2790.
- [12] 文益民, 王耀南, 张莹. 基于可信多数投票的快速概念漂移检测 [J]. 湖南大学学报, 2010, 37(6): 36 – 40.
- [13] 关菁华, 刘大有. 一种挖掘概念漂移数据流的选择性集成算法 [J]. 计算机科学, 2010, 37(1): 204 – 207.
- [14] COVER T M, HART P E. Nearest neighbor pattern classification [J]. IEEE Transactions on Information Theory, 1967, 13(1): 21 – 27.
- [15] YANG Q, WU X. 10 Challenging problems in data mining research [J]. Journal of Information Technology and Decision Making, 2006, 5(4): 597 – 604.
- [16] 鲁婷, 王浩, 姚宏亮. 一种基于中心文档的 KNN 中文文本分类算法 [J]. 计算机工程与应用, 2011, 47(2): 127 – 130.
- [17] AGGARWAL C C, PROCOPIUC C, WOLF J L, et al. Fast algorithm for projected clustering [C]// SIGMOD'99: Proceedings of the 1999 ACM SIGMOD International Conference on Management of Data. New York: ACM Press, 1999: 61 – 71.
- [18] MOISE G, SANDER J, ESTER M. Robust projected clustering [J]. Knowledge Information System, 2008, 14(3): 273 – 398.
- [19] HUANG J Z, NG M K, RONG H, et al. Automated variable weighting in k -means type clustering [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, 27(5): 657 – 668.
- [20] 盛骤, 谢式千, 潘承毅. 概率论与数理统计[M]. 北京: 高等教育出版社, 2006: 241 – 243.
- [21] DEUTEN G, SPENCER L, DOMINGOS P. Mining time-changing data streams [C]// KDD'01: Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2001: 97 – 106.

(上接第 598 页)

持 100 GB 数据的 ETL 操作和数据挖掘算法。而在经典商用数据挖掘工具中, 由于缺乏可扩展性, 一般仅能支持 300 MB 数据的挖掘。

2) 扩展性。

本实验采用 10 节点、30 节点、60 节点、32 节点和 64 节点规模对并行数据处理和并行数据挖掘算法的扩展性进行测试。

图 6(b) 中描述了随着节点数增加的加速比情况。显然, 当节点不变时, 处理的数据量增加, 则加速比接近线性; 当数据量不变时, 增加节点, 加速比接近线性。但是, 当节点数量相对需要处理的数据量过多时, 加速比反而因为非计算开销而远离理论加速比。实验结果表明, 并行数据处理和并行数据挖掘算法具有优秀的扩展能力。

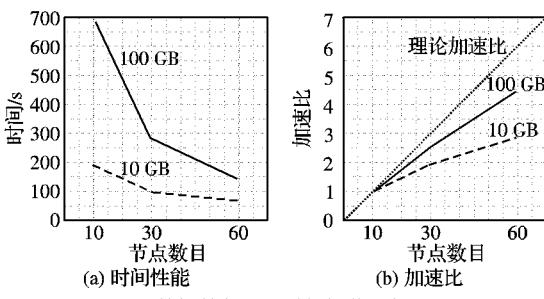


图 6 数据并行处理时间性能和加速比

4 结语

BI-PaaS 搭建于中国移动大云基础设施之上, 以 Hadoop

的强大并行计算和分布存储能力为支撑, 将 ETL、DM、OLAP、Report 等各类 BI 能力并行化, 从而有效地支持电信运营的海量数据分析, 提高电信领域数据分析性能、可扩展性, 降低系统平台成本。

参考文献:

- BURTON B, GEISHECKER L. Organizational structure: business intelligence and information management [R]. Stanford: Gartner Inc, 2006.
- HAN J, KAMBER M. 数据挖掘概念与技术[M]. 范明, 孟小峰, 译. 北京: 机械工业出版社, 2001: 232 – 236.
- 邵凌霜, 李田, 赵俊峰, 等. 一种可扩展的 Web Service QoS 管理框架[J]. 计算机学报, 2008, 31(8): 1458 – 1470.
- 莫则尧, 李晓梅. 工作站网络环境下的并行计算[J]. 计算机学报, 1997, 20(6): 510 – 517.
- 中国移动.“大云”系统[EB/OL].[2011-08-02]. <http://221.183.16.16:8080/>.
- Salesforce 公司. Force. come 应用平台首页[EB/OL].[2011-08-02]. <http://www.salesforce.com/cn/platform/>.
- Google 公司. Google App Engine 应用平台首页[EB/OL].[2011-08-02]. <https://appengine.google.com/>.
- Apache 公司. hadoop 首页[EB/OL].[2011-08-02]. <https://hadoop.apache.org/>.
- DEAN J, GHEMAWAT J. MapReduce: Simplified data processing on large clusters [J]. Communications of the ACM, 2004, 51(1): 107 – 113.
- 吉吉林. 决策树分类技术研究[J]. 计算机工程, 2004, 30(9): 94 – 96.