ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 33,No. 3,May 2018,pp. 416－425
DOI:10. 16337/j. 1004-9037. 2018. 03. 004

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

基于 MapReduce 和上采样的两类非平衡大数据集成分类

翟俊海[1,2]　张明阳[2]　王陈希[3]　刘晓萌[2]　王耀达[2]

(1. 河北省机器学习与计算智能重点实验室,保定,071002; 2. 河北大学数学与信息科学学院,保定,071002; 3. 河北大学计算机科学与技术学院,保定,071002)

摘　要: 提出了一种基于 MapReduce 和上采样的两类非平衡大数据分类方法,该方法分为 5 步:(1)对于每一个正类样例,用 MapReduce 寻找其异类最近临;(2)在两个样例点之间的直线上生成若干个正类样例;(3)以新的正类样例子集的大小为基准,将负类样例随机划分为若干子集;(4)用负类样例子集和正类样例子集构造若干个平衡数据子集;(5)用平衡数据子集训练若干个分类器,并对训练好的分类器进行集成。在 5 个两类非平衡大数据集上与 3 种相关方法进行了实验比较,实验结果表明本文提出的优于这 3 种方法。

关键词: 大数据;非平衡分类;上采样;最近邻

中图分类号: TP181　　**文献标志码**:A

Binary Ensemble Classification for Imbalanced Big Data Based on MapRecuce and Upper Sampling

Zhai Junhai[1,2], Zhang Mingyang[2], Wang Chenxi[3], Liu Xiaomeng[2], Wang Yaoda[2]

(1. Key Lab of Machine Learning and Computational Intelligence, Baoding, 071002, China; 2. College of Mathematics and Information Science, Hebei University, Baoding, 071002, China; 3. College of Computer Science and Technology, Hebei University, Baoding, 071002, China)

Abstract: Based on MapReduce and upper sampling, an approach for imbalanced big data classification is proposed in this paper. The proposed method includes five steps: (1) For each positive instance, its nearest neighbor is found by MapReduce. (2) Some positive instances on the line between the two points are created. (3) According to the cardinality of the set of positive instances, the set of negative instances is partitioned into some subsets. (4) Some balanced subsets are generated with the set of positive instances and the subset of negative instances. (5) Some classifiers are trained by extreme learning machine on the generated balanced subsets, and the trained classifiers are integrated by majority voting for classifying new instances. Experimental comparisons with three related methods are conducted on five imbalanced big data sets. The experimental results show that the proposed method outperforms the three methods.

Key words: big data; imbalanced classification; upper sampling; nearest neighbor

---

基金项目:国家自然科学基金(71371063)资助项目;河北省自然科学基金(F2017201026)资助项目;河北大学自然科学研究计划(799207217071)资助项目。
收稿日期:2016-06-07;**修订日期**:2016-11-29

引 言

随着计算机网络、数据存储、云计算和社会计算等技术的快速发展,数据正以前所未有的速度在不断地增长和积累,大数据处理已经成为学术界和工业界密切关注的问题。大数据是指具有海量(Volume)、多模态(Variety)、变化速度快(Velocity)、蕴含价值高(Value)和可靠性高(Veracity)"5V"特征的数据[1-3]。目前,针对大数据分类的研究主要集中在如何处理大数据量上。解决问题的主流思路包括两种:(1)并行化或分布式方法;(2)基于采样技术的方法。在第一种方法中,由于 MapReduce 编程模型的盛行,大数据分类的并行化或分布式方法基本上都是基于这种编程模型而提出的。例如,Bechini 等利用 MapReduce 编程模型对著名的关联规则挖掘算法 FP-Growth 进行并行化,以实现从大数据中挖掘关联规则[4]。Zhang 等将深度学习和 MapReduce 结合起来,提出了受限波尔兹曼机的分布式学习框架[5],可实现大数据环境中的深度学习。钱宇华等对大数据关联关系度量研究进行了全面的综述[6],具有较高的参考价值。吴启晖等对面向频谱大数据处理的机器学习方法进行了总结,分析了它们各自的特点[7]。吉根林和赵斌综述了时空轨迹大数据模式挖掘与知识发现领域的研究进展[8]。亓峰等对未来大数据环境下的配用电通信网虚拟网络架构及应用进行了研究[9]。第二种方法利用采样技术从大数据集中选择一个子集代替原来的大数据集进行分类。He 等利用不确定性分布,提出了一种从大数据中并行随机采样的方法[10]。与同类算法相比,该方法不仅可以保持原数据超曲面的一致性,而且可以获得非常好的加速比、伸缩比和承载比。针对大数据的 Boosting 集成学习问题,Dubout 等提出了一种自适应采样方法[11]。该方法通过对基本分类器的统计边界行为建模,能够改进大数据 Boosting 集成算法的性能。文献[12]对采样方法研究进行了较全面的综述,具有一定的参考价值。

在现实生活中,很多实际问题中要处理的大数据具有类别非平衡的特点。例如,网络入侵检测、信用卡欺诈检测、恶劣天气预报和医疗诊断等问题。非平衡大数据分类使传统的分类算法面临新的挑战,如何解决非平衡大数据分类问题已成为机器学习和数据挖掘领域的研究热点。处理类别非平衡问题的常用方法大致可分为 4 类[13-15]:(a)数据级的方法,(b)算法级的方法,(c)代价敏感性方法,(d)集成方法。数据级的方法主要利用采样技术,包括对小类样本的随机上采样、对大类样本的随机下采样和基于数据生成的混合采样等。Japkowicz 等提出了基于随机化的上采样和下采样方法[13],并从理论上证明了"在采样之后的数据集合上学习,算法能够获得与原数据集合上等效的学习性能"。Wang 等针对近邻分类器给出了基于特征空间相似性的合成上采样方法 SMOTE[16]。Batista 等提出了基于压缩近邻规则和数据清洗技术的上采样方法[17]。2006 年 Liu 等提出了基于集成策略的独立下采样方法[18]。算法级的方法主要利用归纳偏置、惩罚约束和调整类边界等机制对已有算法(如决策树、支持向量机等)进行改进。代表性的工作包括 Quinlan 提出的通过调整决策树叶结点的概率估计来选择合适的归纳偏置[19];Lin 等提出的对不同类别的样例采用不同惩罚系数的支持向量机分类方法[20]等。代价敏感性方法主要利用样例加权、贝叶斯风险理论等方法设计代价敏感性学习模型。代价敏感性学习的目的是最小化标准数据挖掘或机器学习算法在训练集合上面的错分代价。研究结果表明:通过采用基于代价敏感性方法构建的神经网络[21]、支持向量机[22]和决策树[23]分别可以改善这些传统的数据挖掘和学习算法在非平衡数据集合上的学习性能。集成方法主要包括代价敏感性集成方法和基于数据预处理的集成方法。一般地,代价敏感性集成方法通过在 AdaBoost 算法的权更新公式中引入代价项完成,权更新规则的不同,得到了不同的代价敏感性集成方法。代表性的工作包括 Fan 等提出的 AdaCost 算法[24];Sun 等提出的 AdaCx(x=1,2,3)系列算法[25];Ting 提出的 CSBx(x=1,2)系列算法[26]等。基于数据预处理的集成方法大致又可分为 3 类:基于 Boosting 的方法、基于 Bagging 的方法和混合方法。基于 Boosting

的方法代表性的工作包括 Chawla 等提出的 SMOTEBoost 算法[27]；Seiffert 等提出的 Rusboost 算法[28]等。基于 Bagging 的方法代表性的工作包括 Wang 等提出的 OverBagging 算法和 UnderOverBagging 算法[29]；Barandela 等提出的 UnderBagging 算法[30]等。混合算法代表性的工作包括 Liu 等提出的 EasyEnsemble 算法和 BalanceCascade 算法[31]。

上面这些算法都是针对中小型类别非平衡数据集提出的分类方法，对于类别非平衡的大型数据集，上述算法的效率就会变得非常低，甚至不可行。针对这一问题，在两类分类的框架下，本文提出了一种基于 MapReduce 和上采样的两类非平衡大数据集成分类方法，并在 5 个类别非平衡的大型数据集上进行了实验，实验结果证明本文提出的算法是解决两类非平衡大数据分类问题的一种有效方法。

1 基础知识

本节介绍将要用到的基础知识，包括 MapReduce[32] 和极限学习机(Extreme learning machine, ELM)[33]。ELM 用作分类器对数据进行分类。

1.1 MapRecuce

MapRecuce[32] 是针对大数据处理的一种并行编程框架，它的基本思想包括以下 3 个方面：

(1)MapRecuce 采用分治策略自动地将大数据集划分为若干子集，并将这些子集部署到不同的云计算节点上，并行地对数据子集进行处理；

(2)基于函数编程语言 LISP 的思想，MapRecuce 提供了两个简单易行的并行编程方法：Map 和 Reduce，用它们实现基本的并行计算；

(3)许多系统级的处理细节 MapRecuce 能自动完成，这些细节包括：

(a)计算任务的自动划分和自动部署；

(b)自动分布式存储处理的数据；

(c)处理数据和计算任务的同步；

(d)对中间处理结果数据的自动聚集和重新划分；

(e)云计算节点之间的通讯；

(f)云计算节点之间的负载均衡和性能优化；

(g)云计算节点的失效检查和恢复。

MapRecuce 处理数据的流程如图 1 所示。

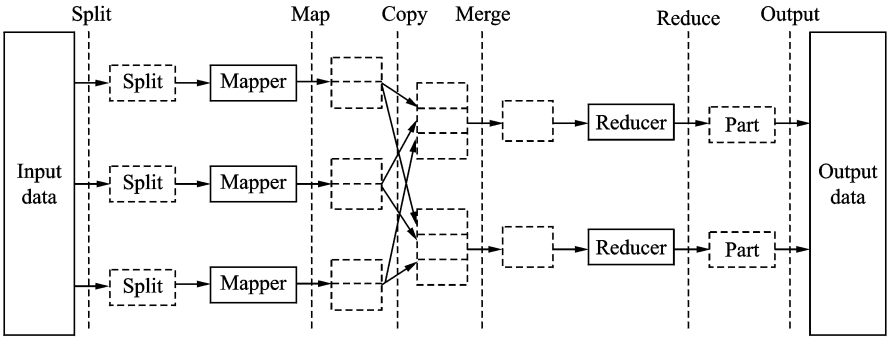


图 1 MapRecuce 处理数据的流程示意图

Fig. 1 Flow chart of data processing by MapRecuce

1.2 极限学习机

ELM[33]是黄广斌等提出的一种训练单隐含层前馈神经网络(如图 2 所示)的简单而有效的算法。ELM 随机生成输入层的权值和隐含层结点的偏置,用分析的方法确定输出层的权值。与其他的单隐含层前馈神经网络训练算法相比,ELM 的优点是不需要迭代调整权参数,具有非常快的学习速度和非常好的泛化能力。而且,黄广斌等证明了 ELM 具有一致逼近能力[34]。

给定训练集 $D=\{(\boldsymbol{x}_i,\boldsymbol{y}_i)\mid \boldsymbol{x}_i\in\mathbf{R}^d,\boldsymbol{y}_i\in\mathbf{R}^k\}$,$1\leqslant i\leqslant n$,具有 m 个隐含层结点的单隐含层前馈神经网络可表示为

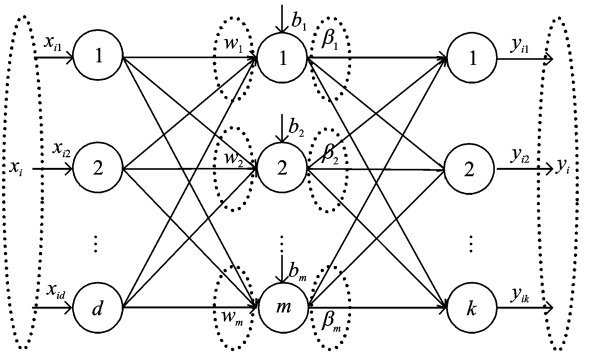


图 2 单隐含层前馈神经网络

Fig. 2 Single-hidden layer feedforward neural network

$$f(x_i)=\sum_{j=1}^{m}\boldsymbol{\beta}_j g(\boldsymbol{w}_j\cdot\boldsymbol{x}_i+b_j)\quad i=1,2,\cdots,n \tag{1}$$

式中:$g(\cdot)$是激活函数;$\boldsymbol{w}_j=(w_{j1},w_{j2},\cdots,w_{jd})^{\mathrm{T}}$ 是输入层结点到隐含层第 j 个结点的权向量;b_j 是隐含层第 j 个结点的偏置,在 ELM 中 w_j 和 b_j 是随机生成的;$\boldsymbol{\beta}_j=(\beta_{j1},\beta_{j2},\cdots,\beta_{jm})^{\mathrm{T}}$ 是隐含层第 j 个结点到输出层结点的权向量,$\boldsymbol{\beta}_j$ 可通过给定的训练集用最小二乘拟合来估计,$\boldsymbol{\beta}_j$ 应满足

$$f(\boldsymbol{x}_i)=\sum_{j=1}^{m}\boldsymbol{\beta}_j g(\boldsymbol{w}_j\cdot\boldsymbol{x}_i+b_j)=\boldsymbol{y}_i\quad i=1,2,\cdots,n \tag{2}$$

式(2)可以写成如下的矩阵形式

$$\boldsymbol{H\beta}=\boldsymbol{Y} \tag{3}$$

其中

$$\boldsymbol{H}=\begin{bmatrix} g(\boldsymbol{w}_1\cdot\boldsymbol{x}_1+b_1) & \cdots & g(\boldsymbol{w}_m\cdot\boldsymbol{x}_1+b_m)\\ \vdots & \vdots & \vdots\\ g(\boldsymbol{w}_1\cdot\boldsymbol{x}_n+b_1) & \cdots & g(\boldsymbol{w}_m\cdot\boldsymbol{x}_n+b_m)\end{bmatrix}$$

$$\boldsymbol{\beta}=[\boldsymbol{\beta}_1^{\mathrm{T}},\boldsymbol{\beta}_2^{\mathrm{T}},\cdots,\boldsymbol{\beta}_m^{\mathrm{T}}]^{\mathrm{T}}$$

$$\boldsymbol{Y}=[\boldsymbol{y}_1^{\mathrm{T}},\boldsymbol{y}_2^{\mathrm{T}},\cdots,\boldsymbol{y}_n^{\mathrm{T}}]^{\mathrm{T}}$$

式中:$\boldsymbol{H}$ 是单隐含层前馈神经网络的隐含层输出矩阵,它的第 j 列是隐含层第 j 个结点相对于输入 $\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n$ 的输出,它的第 i 行是隐含层相对于输入 $\boldsymbol{x}_i$ 的输出。如果单隐含层前馈神经网络的隐含层结点个数等于样例的个数,那么矩阵 $\boldsymbol{H}$ 是可逆方阵。此时,用单隐含层前馈神经网络能零误差逼近训练样例。但一般情况下,单隐含层前馈神经网络的隐含层结点个数远小于训练样例的个数。此时,$\boldsymbol{H}$ 不是一个方阵,线性系统式(3)也没有精确解,但可以通过求解下列优化问题的最小范数最小二乘解来代替式(3)的精确解,即

$$\min_{\boldsymbol{\beta}}=\|\boldsymbol{H\beta}-\boldsymbol{Y}\| \tag{4}$$

上式最小范数最小二乘解可通过下式求得,即

$$\hat{\boldsymbol{\beta}}=\boldsymbol{H}^{+}\boldsymbol{Y}$$

其中 $\boldsymbol{H}^{+}$ 是矩阵 $\boldsymbol{H}$ 的 Moore-Penrose 广义逆矩阵。

极限学习机算法描述如下:

算法 1:极限学习机算法

1. 输入:训练集 $D=\{(\boldsymbol{x}_i,\boldsymbol{y}_i)\mid \boldsymbol{x}_i\in\mathbf{R}^d,\boldsymbol{y}_i\in\mathbf{R}^k,1\leqslant i\leqslant n\}$;激活函数 g;隐含层节点数 m。

2. 输出:权值矩阵 $\hat{\boldsymbol{\beta}}$。

3. for ($j=1$; $i\leqslant m$; $j=j+1$) do

4. 随机给定输入权值 ω_j 和偏置 b_j；
5. end
6. 计算隐含层输出矩阵 $\boldsymbol{H}$；
7. 计算矩阵 $\boldsymbol{H}$ 的广义逆矩阵 $\boldsymbol{H}^+$；
8. 计算权矩阵 $\hat{\boldsymbol{\beta}}=\boldsymbol{H}^+\boldsymbol{Y}$；
9. 输出权矩阵 $\hat{\boldsymbol{\beta}}$。

2 基于 MapReduce 和上采样的两类非平衡大数据集成分类

在两类分类的框架下，讨论类别非平衡大数据集分类问题。一般地，在非平衡学习中，样例数很少的类称为正类，样例数很多的类称为负类。两类非平衡大数据集的大数据量主要体现在负类样例上。算法的基本思想描述如下：首先，对于正类样例集合 S^+ 中的每一个正类样例，在负类样例集合 S^- 中用 MapReduce 寻找其最近邻。然后，在两者之间的连线上上采样若干正类样例点(如图 3 所示)。其次，以上采样后的正类样例集合 S^+ 包含的样例数 $|S^+|=n^+$ 为基准，把负类样例集合随机划分成 $p=\frac{n}{n^+}$ 个子集 $S_1^-, S_2^-, \cdots, S_p^-$。其中，$n$ 为原非平衡大数据集包含的样例数。每个子集 S_i^- $(1\leqslant i\leqslant p)$ 与 S^+ 合并在一起构成一个平衡数据子集 $D_i=S^+\cup S_i^-$ $(1\leqslant i\leqslant p)$。在每一个平衡数据子集 D_i $(1\leqslant i\leqslant p)$ 上，用极限学习机算法训练一个单隐含层前馈神经网络分类器 L_i $(1\leqslant i\leqslant p)$。最后用简单的多数投票法对 p 个分类器进行集成，以分类新的样例。为描述方便，用(Binary ensemble classification for imbalanced big data based on MapReduce and upper sampling, BECIMU)表示提出的算法，算法流程图如图 4 所示。BECIMU 算法的伪码描述如算法 2 所示。

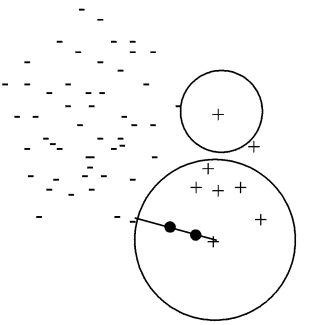


图 3 在正类样例与其负类最近邻的连线上上采样若干正类样例

Fig. 3 Sampling of some points on the line between positive instance and its negative nearest neighbor

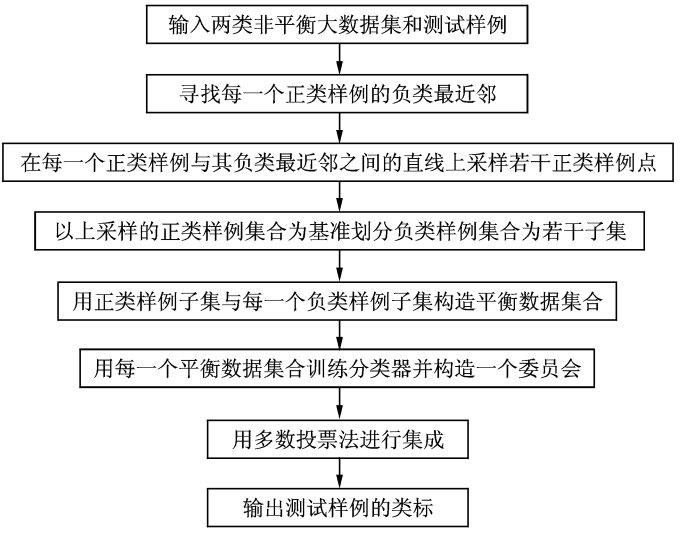


图 4 BECIMU 算法的流程图

Fig. 4 Flow diagram of BECIMU algorithm

算法 2 的第 3～7 步实现正类样例的上采样。其中,第 4 步用 MapReduce 寻找正类样例的异类最近邻,整个算法的计算时间复杂度主要体现在这一步。假定云平台中有 m 个计算节点,显然这一步的计算复杂度为 $O(n)/m$。第 5 步在正类样例与其异类最近邻的连线上上采样,采样点的位置取决于参数 λ,λ 取不同的值可得到不同的采样点。λ 的值越小,上采样点越靠近正类样例点。算法的其他步骤易于理解,不再赘述。

在算法 2 中,MapReduce 的 Map 函数和 Reduce 函数的设计如算法 3 和算法 4 所示。在算法 3 和算法 4 中,$\langle k_1, v_1\rangle$分别是〈起始偏移量,训练样本〉;$\langle k_2, v_2\rangle$分别是〈vector〈欧式距离,训练样本类标志〉,NullWritable〉;$\langle k_3, v_3\rangle$分别是〈测试样本,测试样本的类标志〉。

算法 2: BECIMU 算法

1. 输入:两类非平衡大数据集 $D=S^+\cup S^-$, $|S^+|=n^+$, $|S^-|=n^-$, $n^+\ll n^-$;测试样例 $\boldsymbol{x}$。
2. 输出:$\boldsymbol{x}$ 的类标
3. for ($i=1$; $i\leqslant n^+$; $i=i+1$) do
4. 　对于每一个正类样例 $\boldsymbol{x}_i^+\in S^+$,在 S^- 中用 MapReduce 寻找其异类最近邻 $\boldsymbol{x}_i^-$;
5. 　按公式 $\boldsymbol{x}=\dfrac{\boldsymbol{x}_i^++\lambda\boldsymbol{x}_i^-}{1+\lambda}$,在 $\boldsymbol{x}_i^+$ 和 $\boldsymbol{x}_i^-$ 之间的连线上,上采样若干正类样例点,设 S_i^+ 表示上采样的样例点的集合;
6. 　$S^+=S^+\cup S_i^+$;
7. end
8. 划分负类样例集合 S^- 为 p 个子集 $S_1^-, S_2^-, \cdots, S_P^-$;
9. for ($i=1$; $i\leqslant p$; $i=i+1$) do
10. 　构造 p 个平衡数据子集 $D_i=S^+\cup S_i^-$;
11. 　在 D_i 上,用极限学习机算法训练一个分类器 L_i;
12. end
13. 用多数投票法集成 p 个训练好的分类器 L_i;
14. 用集成系统预测测试样例 $\boldsymbol{x}$ 的类标;
15. 输出 $\boldsymbol{x}$ 的类标。

算法 3:Map 函数

1. 输出:$\langle k_1, v_1\rangle$。
2. 输出:$\langle k_2, v_2\rangle$
3. //遍历所有负类样例 $\boldsymbol{x}_i$,取出其类标志 label;
4. for($i=1$; $i\leqslant n$; $i=i+1$) do
5. 　label-FindLabel ($\boldsymbol{x}_i$);
6. 　//遍历正类样例 $\boldsymbol{x}$,计算其与负类样例之间的欧式距离,并将结果存入 Context;
7. 　for ($\forall\,\boldsymbol{x}\in$ testfile) do
8. 　Distance-EuclideanDistance($\boldsymbol{x}-\boldsymbol{x}_i$);
9. 　Context.write(vector〈Distance,label〉, NullWritable);
10. 　end
11. end
12. 输出$\langle k_2, v_2\rangle$。

算法 4: Reduce 函数

1. 输出:$\langle k_2, v_2\rangle$

2. 输出：〈k_3，v_3〉
3. // 将 vector(Distance，label)添加到 Arraylist 中；
4. ArrayList(Vector〈Distance，label〉)；
5. //对 Arraylist 中所有元素执行排序操作；
6. Sort(ArrayList)
7. //将最近邻添加到 result 中；
8. New ArrayList result；
9. result. add(ArrayList. get(1))；
10. //应用最近邻算法，结果存入 Context 中；
11. Context. write($\boldsymbol{x}$，NN(result))；
12. 输出〈k_3，v_3〉

3 实验结果

为了验证提出的算法的有效性，在 5 个非平衡大数据集上进行了实验，分别与 SMOTE-Vote，SMOTE-Boost 和 SMOTE-Bagging 3 种算法[35]进行了比较。实验所用的云计算平台及各个节点的配置分别列表 1 和表 2 中。

表 1 实验所用云计算平台的配置

Tab. 1 Configuration of cloud computing platform

配置项	配置信息
Operating system	Ubuntu 13.04
Hadoop	Hadoop 0.20.2
JDK	JDK-7u71-linux-i586
Eclipse	Eclipse-java-luna-SR1-linux

表 2 云计算平台节点的配置

Tab. 2 Nodes configuration of the cloud computing platform

配置项	配置信息
CPU	Inter Xeon E5-4603 with two cores，2.0 GZ
内存	8 GB
网卡	Broadcom 5720 QP 1 Gb
硬盘	1 TB

实验所用的 5 个非平衡大数据集分别记为 A，B，C，D 和 E。数据集 A 是由 UCI 数据集 Skin_segment 变换而来，包含 3 679 个正例和 114 039 个负例；数据集 B 由 UCI 数据集 MiniBooNE 变换而来，包含 4 800 个正例和 196 555 个负例；数据集 C 由 UCI 数据集 Cod_rna 变换而来，包含 7 742 个正例和 328 168个负例；数据集 D 是一个人工数据集，包含 150 个正例和 321 191 个负例。数据集 E 是一个 2 类二维服从高斯分布的人工数据集，包含 400 万个样例。其中，正类样例所占比例为 1%。两类服从的高斯分布为

$$p(\boldsymbol{x} \mid \omega_i) \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \qquad i=1,2 \tag{5}$$

其中参数如表 3 所示。

表 3 两个高斯分布的均值向量和协方差矩阵

Tab. 3 Mean vectors and covariance matrices of two Gaussian distributions

i	$\boldsymbol{\mu}_i$	$\boldsymbol{\Sigma}_i$
1	$\begin{pmatrix}1.0\\1.0\end{pmatrix}$	$\begin{pmatrix}0.6 & -0.2\\-0.2 & 0.6\end{pmatrix}$
2	$\begin{pmatrix}2.5\\2.5\end{pmatrix}$	$\begin{pmatrix}0.2 & -0.1\\-0.1 & 0.2\end{pmatrix}$

对于两类非平衡分类问题,设 T 和 F 分别表示实际的正类类标和负类类标,Y 和 N 分别表示预测的正类类标和负类类标,混淆矩阵的定义如图 5 所示。常用的评价两类非平衡分类算法性能的指标有精度(Precision)、召回率(Recall)、几何均值(G-mean)和 F-度量(F-measure),它们的定义如下。

	Y	N
T	TY 真阳性	TN 真阴性
F	FY 假阳性	FN 假阴性

图 5 混淆矩阵

Fig. 5 Confusion matrix

$$\text{Precision}=\frac{\text{TY}}{\text{TY}+\text{TN}} \tag{5}$$

$$\text{Recall}=\frac{\text{TY}}{\text{TY}+\text{FY}} \tag{6}$$

$$\text{G-mean}=\sqrt{\frac{\text{TY}}{\text{TY}+\text{FY}}\times\frac{\text{FN}}{\text{FN}+\text{TN}}} \tag{7}$$

$$\text{F-measure}=\frac{(1+\beta)^2\times\text{Recall}\times\text{Precision}}{\beta^2\times\text{Recall}\times\text{Precision}} \tag{8}$$

其中 β 是一个参数。因为 G-mean 从真阳性率和假阴性率两方面度量了两类非平衡分类算法的性能,所以本文用它作为评价指标。与 SMOTE-Vote,SMOTE-Boost 和 SMOTE-Bagging 三种算法比较的实验结果如表 4 所示。

表 4 本文算法与 3 种算法比较的实验结果

Tab. 4 Experimental results comparison of the proposed algorithm with other three algorithms

算法	A	B	C	D	E
本文算法	0.861 1	0.911 8	0.923 3	0.913 4	0.886 4
SMOTE-Vote	0.835 1	0.901 9	0.884 2	0.910 6	0.852 3
SMOTE-Boost	0.836 4	0.910 0	0.900 0	0.899 7	0.861 9
SMOTE-Bagging	0.852 4	0.902 0	0.911 1	0.913 3	0.843 1

在 MapReduce 框架下,对提出的算法还进行了加速比的比较,即对于相同的数据集在计算节点不同时速度差异,实验结果如表 5 所示。

表 5 加速比的实验结果

Tab. 5 Experimental results of speed up

节点数	A	B	C	D	E
2	670.4	947.2	1 660.4	1 427.4	1 555.1
3	502.1	706.8	1 137.7	1 027.7	1 149.2
4	409.4	578.9	956.9	856.4	989.5
5	375.4	510.1	877.6	744.7	861.7

从表 4 的实验结果可以看出,本文算法的 G-mean 值均高于其他 3 种算法。其原因是 SMOTE 算法仅在同类近邻的连线上采样一个样例点;而本文算法在正类样例与其异类最近邻的连线上采样多个样例点,可以扩大正类样例的学习域。从表 5 的实验结果可以看出,本文算法的加速比也很明显。因此,从这两方面看,本文提出的算法是比较有效的。

4 结束语

针对两类非平衡大数据分类问题,提出了一种基于 MapReduce 和上采样的集成分类算法。该算法利用 MapReduce 的并行计算机制,寻找每一个正类样例的负类最近邻,并在每一个正类样例与其异类

最近邻的连线上采样若干个正类样例点,采样点的个数由用户控制,具有较强的自适应性。另外,本文提出的算法并行计算每一个正类样例到每一个负类样例的距离,极大地降低了计算时间复杂度。在5个数据集上与SMOTE-Vote,SMOTE-Boost和SMOTE-Bagging三种同类方法进行了实验对比,实验结果证明本文提出的方法优于这3种方法。本文提出的算法具有如下两个特点:(1)算法在正类样例与其异类最近邻的连线上采样多个样例点,这样可以扩大正类样例的学习域;(2)算法具有较好的加速比和较高的分类精度。未来进一步的工作包括:(1)在更多、更大的数据集上实验,并对实验结果进行统计分析;(2)将本文提出的算法扩展到多类非平衡问题。

参考文献:

[1] Emani C K, Cullot N, Nicolle C. Understandable big data: A survey[J]. Computer Science Review, 2015,17:70-81.

[2] Zhou Z H, Chawla N V, Jin Y C, et al. Big data opportunities and challenges: Discussions from data analytics perspectives [J]. IEEE Computational Intelligence Magazine, 2014,9(4):62-74.

[3] 孟小峰, 慈祥. 大数据管理:概念、技术与挑战[J]. 计算机研究与发展,2013,50(1):146-169.
Meng Xiaofeng, Ci Xiang. Big data management: Concepts, techniques and challenges [J]. Journal of Computer Research and Development, 2013,50(1):146-169.

[4] Bechini A, Marcelloni F, Segatori A. A MapReduce solution for associative classification of big data [J]. Information Sciences, 2016,332(1):33-55.

[5] Zhang K, Chen X W. Large-scale deep belief nets with MapReduce[J]. IEEE Access, 2014,2(2):395-403.

[6] 钱宇华, 成红红, 梁新彦,等. 大数据关联关系度量研究综述[J]. 数据采集与处理,2015,30(6):1147-1159.
Qian Yuhua, Cheng Honghong, Liang Xinyan, et al. Review for variable association measures in big data[J]. Journal of Data Acquisition and Processing, 2015,30(6):1147-1159.

[7] 吴启晖, 邱俊飞, 丁国如. 面向频谱大数据处理的机器学习方法[J]. 数据采集与处理,2015,30(4):703-713.
Wu Qihui, Qiu Junfei, Ding Guoru. Machine learning methods for big spectrum data processing[J]. Journal of Data Acquisition and Processing, 2015,30(4):703-713.

[8] 吉根林, 赵斌. 时空轨迹大数据模式挖掘研究进展[J]. 数据采集与处理, 2015,30(1):47-58.
Ji Genlin, Zhao Bin. Research progress in pattern mining for big spatiotemporal trajectories[J]. Journal of Data Acquisition and Processing, 2015, 30(1):47-58.

[9] 亓峰, 唐晓璇, 邢宁哲, 等. 未来大数据环境下的配用电通信网虚拟网络架构及应用[J]. 数据采集与处理,2015, 30(3):511-518.
Qi Feng, Tang Xiaoxuan, Xing Ningzhe, et al. Virtual network architecture and application for smart distribution grid in future large data environment[J]. Journal of Data Acquisition and Processing, 2015, 30(3):511-518.

[10] He Q, Wang H, Zhuang F Z, et al. Parallel sampling from big data with uncertainty distribution[J]. Fuzzy Sets & Systems, 2015,258:117-133.

[11] Dubout C, Fleuret F. Adaptive sampling for large scale boosting [J]. Journal of Machine Learning Research, 2014, 15(2): 1431-1453.

[12] 宋寿鹏, 邵勇华, 堵莹. 采样方法研究综述[J]. 数据采集与处理, 2016, 31(3):452-463.
Song Shoupeng, Shao Yonghua, Du Ying. Survey of sampling methods[J]. Journal of Data Acquisition and Processing, 2016,31(3):452-463.

[13] Japkowicz N, Stephen S. The class imbalance problem: A systematic study[J]. Intelligent Data Analysis, 2002,6(5):429-449, 2002.

[14] He H B, Garcia E A. Learning from imbalanced data[J]. IEEE Transactions on Knowledge and Data Engineering, 2009,21(9):1263-1284.

[15] Sun Y M, Wong A K C, Kamel M S. Classification of imbalanced data: A review [J]. International Journal of Pattern Recognition and Artificial Intelligence, 2009,23(4):687-719.

[16] Wang B X, Japkowicz N. Imbalanced data set learning with synthetic samples [C]// IRIS Machine Learning Workshop. Ottawa, Canada: [s. n.], 2004:153-162.

[17] Batista G, Prati R, Monard M. A study of the behavior of several methods for balancing machine learning training data[J]. ACM SIGKDD Explorations Newsletter, 2004,6(1):20-29.

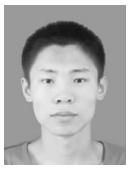
[18] Liu X Y, Wu J, Zhou Z H. Exploratory under sampling for class imbalance learning[C]// Proceedings of the 2006 International Conference on Data Mining. Las Vegas, Nevada, USA: [s. n.], 2006:965-969.

[19] Quinlan J R. Improved estimates for the accuracy of small disjuncts [J]. Machine Learning, 1991, 6:93-98.

[20] Lin Y, Lee Y, Wahba G. Support vector machines for classification in nonstandard situations [J]. Machine Learning, 2002, 46:191-202.
[21] Zhou Z H, Liu X Y. Training cost-sensitive neural networks with methods addressing the class imbalance problem[J]. IEEE Transactions on Knowledge and Data Engineering, 2006,18(1):63-77.
[22] Batuwita R, Palade V. FSVM-CIL: Fuzzy support vector machines for class imbalance learning[J]. IEEE Transactions on Fuzzy Systems, 2010,18(3):558-571.
[23] Ting K M. An instance-weighting method to induce cost-sensitive trees[J]. IEEE Transactions on Knowledge and Data Engineering, 2002,14(3):659-665.
[24] Fan W, Stolfo S J, Zhang J, et al. Adacost: Misclassification cost-sensitive boosting[C]// the 6th Int Conf Mach Learning. San Francisco, CA:[s. n.], 1999: 97-105.
[25] Sun Y, Kamel M S, Wong A AK, et al. Cost-sensitive boosting for classification of imbalanced data [J]. Pattern Recognition, 2007,40(12):3358-3378.
[26] Ting K M. A comparative study of cost-sensitive boosting algorithms[C]// Proc 17th Int Conf Mach Learning. Stanford, CA: [s. n.], 2000:983-990.
[27] Chawla N V, Lazarevic A, Hall L O, et al. SMOTEBoost: Improving prediction of the minority class in boosting[C]// Proceedings of the 2003 European Conference on Principles and Practice of Knowledge Discovery in Databases. Cavtat Dubrovnik, Croatia: [s. n.], 2003:107-119.
[28] Seiffert C, Khoshgoftaar T, Hulse J V, et al. Rusboost: A hybrid approach to alleviating class imbalance [J]. IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans, 2010,40(1):185-197.
[29] Wang S, Yao X. Diversity analysis on imbalanced data sets by using ensemble models[C]// IEEE Symp Comput Intell Data Mining. Nashville, Tennessee: IEEE, 2009:324-331.
[30] Barandela R, Valdovinos R M, Sanchez J S. New applications of ensembles of classifiers[J]. Pattern Analysis & Applications, 2003,6:245-256.
[31] Liu X Y, Wu J, Zhou Z H. Exploratory undersampling for classimbalance learning[J]. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, 2009,39(2):539-550.
[32] Dean J, Ghemawat S. MapReduce: Simplified data processing on large clusters[J]. Communications of the ACM, 2008,51(1):107-113.
[33] Huang G B, Zhu Q Y, Siew C K. Extreme learning machine: Theory and applications[J]. Neurocomputing, 2006, 70(1-3): 489-501.
[34] Huang G B, Chen L, Siew C K. Universal approximation using incremental constructive feedforward networks with random hidden nodes[J]. IEEE Transactions on Neural Networks, 2006,17(4):879-892.
[35] Chawla N V, Bowyer K W, Hall L O, et al. SMOTE: Synthetic minority over-sampling technique [J]. Journal Artificial Intelligence Research, 2002, 16:321-357.

作者简介:

翟俊海(1964-),男,教授,研究方向:机器学习与数据挖掘,E-mail:mczjh@126.com。



张明阳(1991-),男,硕士研究生,研究方向:云计算与大数据处理。

王陈希(1988-),男,硕士研究生,研究方向:机器学习。

刘晓萌(1987-),女,硕士研究生,研究方向:机器学习。

(编辑:刘彦东)