DOI:10.13718/j.cnki.xsxb.2021.07.013

基于边界条件 GAN 的不平衡大数据模糊分类①

杨 琳[1], 徐慧英[2], 马文龙[1]

1. 衢州职业技术学院 信息工程学院, 浙江 衢州 324000; 2. 浙江师范大学 数学与计算机科学学院, 浙江 金华 321004

摘要: 针对大数据分类中的不平衡问题, 本文提出一种基于边界条件生成式对抗网络(Boundary Conditional Generative Adversarial Networks, BCGAN)的不平衡大数据模糊分类算法, 通过在多数类数据和少数类数据的决策边界附近引入一个边界少数类到过样本, 生成更合适的少数类数据来提高分类性能. 将处理过的平衡数据转换成概率索引表, 数据和属性分别以行和列的形式呈现, 计算每个数据属性中存在的唯一符号的隶属度, 然后设计相关模糊朴素贝叶斯(Correlative Fuzzy Naive Bayes, CFNB)分类器进行数据分类. 本文给出 MapReduce 框架下大数据模糊分类的并行实现. 实验结果表明: 所提基于 BCGAN 的不平衡大数据模糊分类准确度优于其他现有方法, 说明该方法具有可行性和有效性.

关 键 词: 大数据; 不平衡; 边界条件生成式对抗网络; 相关模糊朴素贝叶斯

中图分类号: TP393 **文献标志码**: A **文章编号**: 1000-5471(2021)07-0097-06

在大数据时代, 数据已成为一种新的战略资源, 是推动创新的重要因素, 并且正在改变各个领域研究的方式以及人们的生活方式和思维方式[1], 许多国家相继发布了一系列大数据技术计划, 大力推动了大数据的研究和应用[2-3]. 目前的研究一直致力于识别和分析每一个领域的巨大数据, 大数据应用领域有医疗服务、银行业、市场营销等[4]. Cheng 等[5]研究了大数据挖掘技术在智能生产中的应用.

大数据分类是识别输入大数据所属的类的过程, 分类最流行的方法之一是通过使用给定的数据集训练机器学习算法来构造分类模型[6], Varatharajan 等[7]提出支持向量机的心电图大数据分类方法, 将支持向量机(Support Vector Machine, SVM)模型与加权核函数方法结合使用, 从输入的心电图信号中分类更多特征. Lakshmanaprabu 等[8]使用随机森林分类器, 开发了基于物联网(IoT)的医疗系统大数据分析, 使用改进的蜻蜓算法(Improved Dragonfly Algorithm, IDA)从数据库中选择最佳属性获得更好的分类. 张龙翔等[9]提出面向分布式数据流大数据分类的多变量决策树, 设计了几何轮廓相似度的多变量决策树用于大数据分类. 但是上述方法都没有考虑到不平衡数据的处理问题, 如果数据集不平衡, 机器学习等分类算法不能够正确学习少数类数据, 倾向于占数据集很大比例的大多数类, 这可能会导致分类结果有偏差和决策错误[10].

为了减轻类不平衡问题, 通常使用数据采样技术通过任一类中的数据数量来调整不平衡数据. 根据调整的类别, 可以将它们分为欠采样技术和过采样技术[11]. 欠采样在多数类中删除数据, 直到其数目等于少数类中的数据数为止, 欠采样技术由于平衡数据删除而遭受信息丢失的问题. 过采样技术为少数类生成数据与多数类平衡, 常用的过采样方法包括合成少数类过采样技术(Synthetic Minority Over-sampling Technique, SMOTE)、自适应合成采样(Adaptive Synthetic Sampling, ADASYN)和边界 SMOTE[12]. 但是, 过

① 收稿日期: 2020-07-04
基金项目: 浙江省自然科学基金项目(LY15E050007), 浙江省高等学校访问工程师校企合作项目(FG2017139); 衢州市科学技术指导性项目(2018007).
作者简介: 杨 琳, 硕士, 讲师, 主要从事计算机应用研究.

采样方法存在分类模型会被过度拟合到训练数据的问题. 另外，基于SMOTE的数据合成方法有时会产生多数类数据而不是少数类数据.

在研究现有大数据分类方法和采用方法的基础上，本文提出一种基于边界条件生成式对抗网络(Generative Adversarial Networks，GAN)的不平衡大数据分类方法，该方法利用条件GAN的类信息来产生少数类特征数据，然后在数据决策边界引入边界少数类到过样本，生成合适的少数类数据来提高分类性能. 基于相关因子和模糊理论，本文设计了相关模糊朴素贝叶斯分类方法对平衡大数据进行分类，并给出MapReduce框架下大数据分类的并行实现.

1 边界条件GAN

BCGAN利用决策边界条件GAN的类信息生成合适的少数类数据来提高分类性能. GAN由一个生成器和一个鉴别器组成. 条件GAN的结构和基本学习方法与GAN相似，区别是条件GAN的发生器和鉴别器考虑给定的条件. 为了提高分类准确性，本文搜索位于决策边界附近的少数类数据，将其与其他少数类数据区分开. 使用borderline-SMOTE的边界样本选择方法找到边界少数类数据，步骤为：对于少数类中的每个数据实例，使用k最近邻(k-Nearest Neighbor，k-NN)算法计算其k个最近的数据实例，并得出其子集. 对于每个子集，如果属于多数类数据样本的数量大于或等于子集的大小，则将子集中的少数类数据视为边界少数类数据. 由于数据样本距离决策边界较远，因此将其保留在少数类别中. 最终，边界少数类包含原始少数类中靠近决策边界的数据.

BCGAN的目标是沿着多数类和少数类之间的决策边界生成少数类数据，需要对BCGAN进行训练. ①为给定的多数类和少数类计算边界少数类；②将类别信息以及来自高斯分布的随机选择的噪声输入发生器，生成器根据给定的输入数据生成伪造数据；③鉴别器试图通过使用类信息来区分真实数据和生成的数据. 根据条件GAN的损失函数，生成器和鉴别器会更新参数，并且通过重复此过程使损失最小化. BCGAN生成器可以生成反映边缘少数群体特征的数据.

训练完BCGAN之后，基于噪声和边界少数类数据，生成器可以产生与实际边界少数类数据相似的少数类数据，直到多数和少数类数据相同. 此时，这两个类具有相同的数据大小，即得出平衡的数据. 将生成的数据与现有的训练数据进行组合，然后将它们用于训练分类器.

2 相关模糊朴素贝叶斯大数据分类算法

2.1 相关模糊朴素贝叶斯

现有的模糊朴素贝叶斯(Fuzzy Naive Bayes，FNB)分类器利用朴素贝叶斯(Naive Bayes，NB)和基于模糊的方法进行数据分类. 本文设计了相关模糊朴素贝叶斯(Correlative Fuzzy Naive Bayes，在CFNB)分类方法. CFNB分类器中，作为输入的训练数据需要表示为概率索引表，概率索引表表示数据样本作为数据矩阵，概率索引表的行和列表示数据和它们各自的属性. 本文所提出的CFNB分类器的训练样本表示为

$$T = T_{p,q},\ 1 \leqslant p \leqslant d,\ 1 \leqslant q \leqslant a \tag{1}$$

其中T表示分类器的训练样本，$T_{p,q}$表示概率索引表第q属性中的p数据样本. d和a分别表示训练数据集中存在的总数据样本和属性. 式(2)给出向量形式的类.

$$G = \{g_p,\ 1 \leqslant p \leqslant d\} \tag{2}$$

其中G表示向量的类，g_p表示p个数据样本的类. 数据样本中存在的属性对数据分类有更大的贡献. 考虑具有多个属性的训练数据样本. 因此，数据样本的属性表示为

$$H = \{h_q,\ 1 \leqslant q \leqslant a\} \tag{3}$$

其中H表示数据样本的属性，h_q表示数据样本的第q个属性，在每个属性下分类的数据样本具有唯一的符号，CFNB分类器根据属性内的唯一数据符号计算模糊隶属度. 对于CFNB分类器隶属度的计算，考

虑训练样本中唯一符号的第 q 个属性，S 表示训练样本中的符号数量，符号的第 q 个属性由 $h_q \in m^s$ 表示，s 表示符号变量，取值范围为 $1 < s < S$，CFNB 分类器提供的训练样本第 q 个属性中 s 符号的隶属度可表示为 $u_q^s = |m_q^s|/d$，其中 $|m_q^s|$ 表示第 q 个属性中 s 符号的总出现次数，d 表示属性中的数据样本数量. CFNB 分类器将数据样本分类为 K 个类别，类总数的变化表示为 G_k，其中 k 表示类别变量，取值范围为 $1 < k < K$. CFNB 分类器还为地面真相信息计算每个类的隶属度，带有基本事实信息第 k 类的隶属度可表示为 $u_c^k = |m^k|/d$，其中 c 表示类别.

CFNB 分类器与 FNB 不同之处：为训练数据集中存在的每个属性找到虚拟相关因子. 式(4) 表示训练数据每个属性之间的虚拟相关性.

$$C^k = f(h_1, h_2, \cdots, h_q, \cdots, h_a) \tag{4}$$

其中 q 表示属性变量，a 表示训练数据集中存在的总属性数，C^k 表示第 k 类中属性的虚拟相关性，$f(\cdot)$ 表示相关函数. 通过将属性和训练样本的符号表示为对角矩阵来构造数据样本属性之间的相关函数. 式(5) 表示训练数据属性之间的相关函数.

$$f(h_1, h_2, \cdots, h_q, \cdots, h_a) = \frac{1}{1+2+\cdots+(a-1)} \sum_{q=1}^{a} \sum_{s=q+1}^{a} r(h_s, h_q) = \frac{2}{a(a-1)} \sum_{q=1}^{a} \sum_{s=q+1}^{a} r(h_s, h_q) \tag{5}$$

其中 h_q 表示数据样本的第 q 个属性，h_s 表示数据样本的第 s 个属性，$r(h_s, h_q)$ 表示第 s 个和第 q 个属性之间的相关性. CFNB 分类器考虑相关因子，找到训练数据中存在的数据样本之间的关系，并找到训练集中存在的独立数据样本的相关性. 查找属性中存在的唯一符号之间关系的相关因子表示为

$$r(h_s, h_q) = \frac{correlativer(h_s, h_q) + 1}{2} \tag{6}$$

其中函数 $correlativer(h_s, h_q)$ 表示皮尔逊的相关系数，可以查找数据样本之间的线性相关性. 皮尔逊相关系数的一般表达式为

$$correlativer(h_s, h_q) = \frac{\sum_{p=1}^{d} (t_{pq} - \bar{t}_q)(t_{ps} - \bar{t}_s)}{\sqrt{\sum_{p=1}^{d} (t_{pq} - \bar{t}_q)^2} \sqrt{\sum_{p=1}^{d} (t_{ps} - \bar{t}_s)^2}} \tag{7}$$

其中 d 表示属性中的数据样本数量，p 为数据样本变量，取值范围为 $[1, d]$，t_{pq} 表示第 q 属性中存在的数据样本，$\bar{t}_q$ 表示第 q 属性中存在的数据样本的平均值，t_{ps} 表示第 q 属性中的唯一数据符号，$\bar{t}_s$ 表示第 q 属性中唯一数据符号的平均值. CFNB 分类器训练的最终输出包含来自属性的隶属度、来自地面真相信息的隶属度和相关因子. CFNB 分类器的输出表示为

$$CFNB = \{u_q^k, u_c^k, c^k\} \tag{8}$$

其中 u_q^k 表示第 q 个属性第 k 类的隶属度. 属性的隶属度具有 $(d \times S)$ 的大小，而地面真值信息的隶属度具有 $(1 \times K)$ 的大小. 其中，d 表示属性中的数据样本数量，S 表示训练样本中的符号数量，K 表示数据样本分类类别数量. 每个类表示属性唯一符号之间的相关系数具有 $(1 \times K)$ 的大小，CFNB 分类器的训练结果具有 $(d \times S + 2K)$ 的总大小.

2.2　CFNB 的 MapReduce 并行实现

MapReduce 框架由 mapper 和允许大型数据集同时运行的 reducer 组成，本文给出 CFNB 分类器训练和测试阶段的大数据分类过程. 在训练阶段，训练数据被输入到 MapReduce 框架，CFNB 在训练阶段的 MapReduce 框架如图 1 所示.

CFNB 在测试阶段的 MapReduce 框架如图 2 所示.

提供给 mapper 的测试数据表示为 X，对测试数据 X 进行分区，测试数据包含 U 个部分数据样本和多个

属性，计算每个映射器的隶属度、相关函数和 mapper 数据数量. 最后，mapper 将信息提供给 reducer 合并信息，其中 k 小于 U，并提供有关测试数据样本各部分类变量的信息，并最终分类.

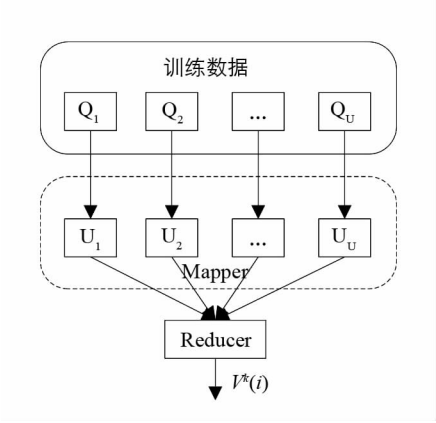

图 1 训练阶段 CFNB 的 MapReduce 框架

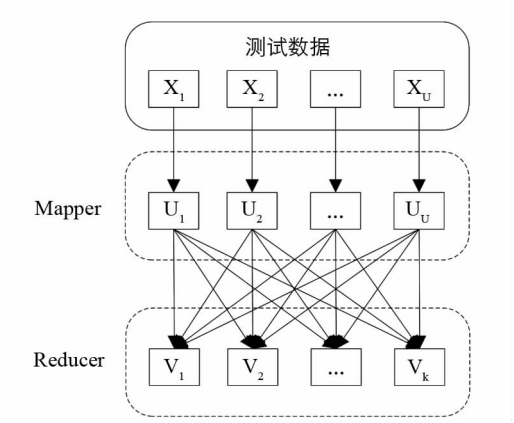

图 2 测试阶段 CFNB 的 MapReduce 框架

3 实验结果与分析

为验证本文所提算法的性能，将该算法和现有其他算法进行实验，所有实验在个人计算机中的 Java 平台上完成. 计算机配置如下：Windows 10 操作系统、4 GB RAM 和英特尔 I7 处理器. 实验大数据集是著名的 PokHand 数据集的不平衡版，由超过 100 万个具有 10 个预测属性的训练和测试样本组成.

对比算法有：文献[13]中基于混合采样策略的改进随机森林不平衡数据分类算法；文献[14]中基于鲸鱼优化＋ SMOTE＋双向递归神经网络的大数据分类算法；文献[15]中基于加权极限机器学习(WELM)和 PSO 的大数据分类算法. 表 1 给出了不同算法的分类准确度对比结果.

表 1 不平衡大数据分类准确度

数据集	文献[13]	文献[14]	文献[15]	本文
Poker_0－2	81.51	84.89	83.67	98.56
Poker_0－3	92.1	94.06	93.72	96.25
Poker_0－4	95.45	97.75	96.16	99.2
Poker_0－5	95.21	97.23	96.54	99.7
Poker_0－6	94.45	97.44	95.74	99.69
Poker_0－7	87.98	89.19	85.67	99.95
Poker_1－2	80.81	82	81.72	91.13
Poker_1－3	92.56	94.61	93.61	95.21
Poker_1－4	94.2	96.77	94.54	99.07
Poker_1－5	89.98	92.4	90.67	99.58
Poker_1－6	93.15	97.47	94.49	99.78
Poker_1－7	94.03	93.15	91.34	99.96

从表 1 可以看出，本文算法在不平衡大数据条件下分类准确度都高于其他 3 种方法，这是因为本文算法首先使用基于边界条件的 GAN 对不平衡数据进行处理，得到平衡数据，提高了分类准确度. 另外，使用虚拟相关因子的模糊 NB 对数据进行分类，进一步提高了分类准确率. 文献[14]中基于鲸鱼优化＋SMOTE＋双向递归神经网络的大数据分类算法，分类准确度仅次于本文算法，这是因为使用鲸鱼优化＋SMOTE 对不平衡数据进行了处理，然后使用深度学习的双向递归神经网络提高了分类准确度. 文献[13]算法对不平衡大数据集表现出较好的分类结果，这是因为该算法使用混合采样方法对平衡数据进行了处理，为了验证本文算法的时间性能，得到平衡数据，提高了分类准确度.

针对加利福尼亚大学尔湾分校(UCI)机器存储库的定位数据集进行时间性能验证，该数据集包含人活动的信息. 且包含 8 个属性下 164 860 个样本实例. 大数据分类时间对比结果如图 3 所示.

由图 3 可以看出，本文算法的大数据分类时间最少，这是因为在 MapReduce 框架进行本文算法的并行化实现，大大减少了分类时间. 虽然文献[14]中方法分类准确度较高，但是递归神经网络耗时较大，导致分类时间最多.

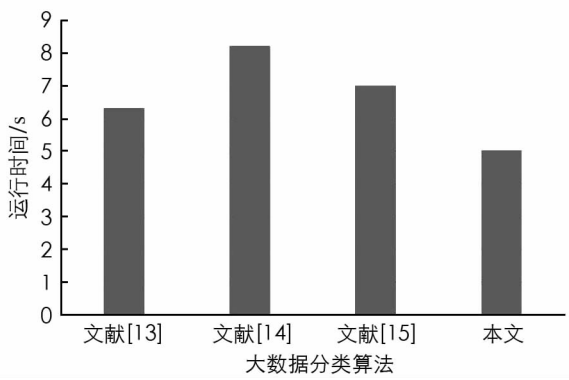

图 3　不同算法的运行时间对比

4　结　　语

本文提出一种基于 BCGAN 的不平衡大数据模糊分类算法，该算法使用 BCGAN 在多数类数据和少数类数据的决策边界附近引入一个边界少数类到过样本，生成更合适的少数类数据来提高分类性能，处理不平衡大数据，得到利于分类的平衡大数据，然后设计了基于相互因子和模糊理论的 CFNB 分类器. 将得到的平衡数据转换成概率索引表，通过相互因子和隶属度的引入进一步提高大数据分类性能，最后给出了 MapReduce 框架下的并行实现，降低了分类时间. 实验结果表明，与现有其他方法比较，针对不平衡率数据集，本文算法具有最优的分类准确度和最低的分类时间，说明该方法具有可行性和有效性.

参考文献：

[1] GHANI N A, HAMID S, TARGIO HASHEM I A, et al. Social Media Big Data Analytics: a Survey [J]. Computers in Human Behavior, 2019, 101: 417-428.

[2] 姜丽丽，李叶飞，豆龙龙，等. 面向大数据的图模式挖掘概率算法 [J]. 计算机应用研究，2020，37(12)：3545-3551.

[3] GARCÍA-GIL D, LUENGO J, GARCÍA S, et al. Enabling Smart Data: Noise Filtering in Big Data Classification [J]. Information Sciences, 2019, 479: 135-152.

[4] WANG Y C, KUNG L, BYRD T A. Big Data Analytics: Understanding Its Capabilities and Potential Benefits for Healthcare Organizations [J]. Technological Forecasting and Social Change, 2018, 126: 3-13.

[5] CHENG Y, CHEN K, SUN H M, et al. Data and Knowledge Mining with Big Data towards Smart Production [J]. Journal of Industrial Information Integration, 2018, 9: 1-13.

[6] LUECHTEFELD T, MARSH D, ROWLANDS C, et al. Machine Learning of Toxicological Big Data Enables Read-across Structure Activity Relationships (RASAR) Outperforming Animal Test Reproducibility [J]. Toxicological Sciences, 2018, 165(1): 198-212.

[7] VARATHARAJAN R, MANOGARAN G, PRIYAN M K. A Big Data Classification Approach Using LDA with an Enhanced SVM Method for ECG Signals in Cloud Computing [J]. Multimedia Tools and Applications, 2018, 77(8): 10195-10215.

[8] LAKSHMANAPRABU S K, SHANKAR K, ILAYARAJA M, et al. Random Forest for Big Data Classification in the Internet of Things Using Optimal Features [J]. International Journal of Machine Learning and Cybernetics, 2019, 10(10): 2609-2618.

[9] 张龙翔，曹云鹏，王海峰. 面向大数据复杂应用的 GPU 协同计算模型 [J]. 计算机应用研究，2020，37(7)：2049-2053.

[10] CARVALHO A M D, PRATI R C. Improving kNN Classification under Unbalanced Data. a New Geometric Oversampling Approach [C]//2018 International Joint Conference on Neural Networks (IJCNN). Rio de Janeiro: IEEE, 2018.

[11] HASANIN T, KHOSHGOFTAAR T M, LEEVY J, et al. Investigating Random Undersampling and Feature Selection on Bioinformatics Big Data [C]//2019 IEEE Fifth International Conference on Big Data Computing Service and Applications (BigDataService). Newark: IEEE, 2019.

[12] POLAT K. A Hybrid Approach to Parkinson Disease Classification Using Speech Signal: The Combination of SMOTE and Random Forests [C]//2019 Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT). Istanbul: IEEE, 2019.

[13] 郑建华, 刘双印, 贺超波, 等. 基于混合采样策略的改进随机森林不平衡数据分类算法 [J]. 重庆理工大学学报(自然科学), 2019, 33(7): 113-123.

[14] HASSIB E M, EL-DESOUKY A I, LABIB L M, et al. WOA+BRNN: an Imbalanced Big Data Classification Framework Using Whale Optimization and Deep Neural Network [J]. Soft Computing, 2020, 24(8): 5573-5592.

[15] UTOMO O K, SURANTHA N, ISA S M, et al. Automatic Sleep Stage Classification Using Weighted ELM and PSO on Imbalanced Data from Single Lead ECG [J]. Procedia Computer Science, 2019, 157: 321-328.

Fuzzy Classification of Unbalanced Big Data Based on Boundary Condition GAN

YANG Lin[1], XU Hui-ying[2], MA Wen-long[1]

1. *School of Information Engineering, Quzhou College of Technology, Quzhou Zhejiang 324000, China;*

2. *College of Mathematics and Computer Science, Zhejiang Normal University, Jinhua Zhejiang 321004, China*

Abstract: Aiming at the imbalance problem in big data classification, an unbalanced big data fuzzy classification algorithm based on boundary condition generative adversarial networks (BCGAN) has been proposed. In this method, BCGAN oversampling method is proposed by introducing a boundary minority class to oversampling near the decision boundary of majority class data and minority class data, generating more appropriate minority class data to improve the classification performance. The processed balance data is transformed into probability index table, and the data and attributes are presented in the form of row and column respectively. The membership degree of the unique symbol in each data attribute is calculated, and then the data category is obtained by means of the correlative fuzzy naive Bayes (CFNB) classifier. Then, the parallel implementation of big data fuzzy classification in MapReduce framework is given. The experimental results show that the accuracy of the proposed method is better than that of other existing methods, indicating the feasibility and effectiveness of the proposed method.

Key words: big data; imbalance; boundary condition generative adversarial network; correlative fuzzy naive bays

责任编辑 夏 娟