DOI:10.13718/j.cnki.xsxb.2021.03.001

基于差异灰狼优化决策树的大数据分类方法①

吕廷勤, 魏 萌

郑州师范学院 信息科学与技术学院, 郑州 450044

摘要: 针对现有大数据分类算法中存在准确率低的问题, 本文提出一种基于差异灰狼优化决策树的大数据分类方法. 该方法首先将复杂的大数据输入 Map-Reduce 框架中, 采用主成分分析法对输入数据进行降维; 然后利用支持向量机对压缩后的数据进行粗略分类; 最后采用基于差异灰狼优化的决策树对支持向量机输出的类标签进行精细分类, 获得更高的分类准确度. 实验结果表明, 相比于其他分类算法, 本文提出的方法在复杂大数据分类方面具有明显的优势.

关 键 词: 大数据; 数据分类; 差异灰狼优化; 决策树

中图分类号: TP391 **文献标志码**: A **文章编号**: 1000-5471(2021)03-0001-06

随着信息技术在医疗服务方面的飞速发展, 整个医疗系统采集到了海量数据信息. 但是, 由于医疗信息数据的复杂性, 造成信息利用率相对较低. 医疗数据通常是信息化程度非常高的数据类型, 包含了大量的小样本数据和非结构化数据, 从海量医疗数据中有效挖掘蕴含其中的价值信息, 提高数据利用率已成为当前医疗行业内的热点问题[1-2].

数据挖掘是从大量数据中通过统计学、人工智能、机器学习等方法搜索隐藏于其中的信息的过程[3]. 数据挖掘通过寻找未知数据的模式与规律, 解决大数据中分类、聚类、关联和预测等问题[4]. 目前, 已有很多基于数据挖掘对数据进行分类处理的算法. 文献[5]提出了一种基于机器学习的同步模型, 该模型通过支持向量机和神经网络从真实的医学数据集中提取临床特征对高脂血症进行分类, 实现了提高效率和降低工作强度的目的. 文献[6]将统计技术与加权 K 最邻近(K-Nearest Neighbor, K-NN)分类器结合使用, 对公共微阵列基因组数据的分类效果明显, 但是 K-NN 模型存在不容易解释, 且计算量大的问题. 文献[7]提出了一种基于布谷鸟搜索算法优化 Gabor 滤波器组的方法, 基于大数据技术对乳腺钼靶进行分类, 提高数据处理的效率和准确度. 但是该类模型需要调整参数, 存在过度拟合问题. 文献[8]利用语言神经模糊特征提取模型对医学数据进行分类, 该模型使用语言模糊化过程生成处理不确定性问题的隶属度值, 然后在神经模糊模型中混合特征提取算法来简化显著贡献的特征, 最后将简化特征输入人工神经网络进行分类. 神经模糊模型能够有效进行数据分类, 解决大数据中存在的不平衡问题, 但是生成模糊规则复杂性高, 并且全局建模的稳定性差[9].

针对上述医疗大数据分类算法中存在的一些局限性, 本文提出了一种基于差异灰狼优化决策树的大数据分类方法, 该方法首先将复杂的大数据输入 MapReduce 框架中, 通过主成分分析(Principle Component Analysis, PCA)操作来减少数据量, 其次利用支持向量机(Support Vector Machine, SVM)对精简后的数据进

① 收稿日期: 2020-02-11
基金项目: 河南省教育厅青年基金项目(QN2016182).
作者简介: 吕廷勤, 硕士, 讲师, 主要从事计算机应用研究.

行分类，最后利用基于差异灰狼优化的决策树对SVM输出的类进行精细分类，获得更高的分类准确度.

1 灰狼优化算法

灰狼优化(Gray Wolf Optimization，GWO)[10]是一种启发式群智能优化算法，该算法根据自然界中灰狼的等级制度与狩猎行为研究而成. 在灰狼优化模型中，狼群社会分为4个等级：领导者α负责狩猎、休息等活动决策，从属者β辅助α在活动中做出决策，执行者δ跟随领导者和从属者，并执行两者的决策，跟随者ω等级最低，服从α,β和δ三者的支配. 在灰狼优化算法寻优的过程中，最优解、次优解及第三解分别为α,β和δ，ω对应其余解.

狼群的捕食过程分为跟踪、追捕和攻击3个阶段. 在狩猎过程中，灰狼跟踪猎物的行为在数学上可以定义为

$$\vec{I} = \vec{G} \cdot \vec{Z}_b(x) - \vec{Z}(x) \tag{1}$$

$$\vec{Z}(x+1) = \vec{Z}_b(x) - \vec{F} \cdot \vec{I} \tag{2}$$

式(1)、式(2)中：x表示当前迭代次数，$\vec{Z}_b$和$\vec{Z}$分别表示猎物和灰狼的位置向量，$\vec{I}$表示猎物与灰狼之间的距离. $\vec{G}$和$\vec{F}$表示协同系数向量，可以定义为

$$\vec{F} = 2\vec{e} \cdot \vec{h}_1 - \vec{e} \tag{3}$$

$$\vec{G} = 2 \cdot \vec{h}_2 \tag{4}$$

式(3)、式(4)中：$\vec{h}_1$和$\vec{h}_2$是一维随机向量，在[0，1]内取值，$\vec{e}$为系数向量，元素随迭代次数的增加从2至0线性递减.

追捕阶段是灰狼通过识别猎物位置并开展追捕包围猎物的过程，在此过程中领导者α制定策略，从属者β和执行者δ做出配合. 但是，在迭代过程中对最优解位置是未知的. 因此，在建立数学模型时，需要考虑α、β和δ共3个潜在解的位置

$$\begin{aligned}
\vec{I}_\alpha &= \vec{G}_1 \cdot \vec{Z}_\alpha - \vec{Z} \\
\vec{I}_\beta &= \vec{G}_2 \cdot \vec{Z}_\beta - \vec{Z} \\
\vec{I}_\delta &= \vec{G}_3 \cdot \vec{Z}_\delta - \vec{Z}
\end{aligned} \tag{5}$$

$$\begin{aligned}
\vec{Z}_1 &= \vec{Z}_\alpha - \vec{F}_1 \cdot \vec{I}_\alpha \\
\vec{Z}_2 &= \vec{Z}_\beta - \vec{F}_2 \cdot \vec{I}_\beta \\
\vec{Z}_3 &= \vec{Z}_\delta - \vec{F}_3 \cdot \vec{I}_\delta
\end{aligned} \tag{6}$$

$$\vec{Z}(x+1) = \frac{\vec{Z}_1 + \vec{Z}_2 + \vec{Z}_3}{3} \tag{7}$$

攻击是捕食过程中的最后阶段，灰狼通过攻击猎物结束狩猎，获得最优解. 从数学上讲，降低$\vec{e}$值来表示接近猎物，$\vec{e}$值的递减也缩小了波动区间$F \in [-e, e]$. 在$\vec{e}$值从2降至0的过程中，当$|F| \leqslant 1$时表明狼群在接近猎物，反之$|F| > 1$表示狼群远离猎物，导致算法陷入局部最优，失去最优解位置. $\vec{e}$值的更新公式可以表示为

$$e = 2 - 2x/Iter_{\max} \tag{8}$$

式(8)中：$Iter_{\max}$表示最大迭代次数.

2　基于差异灰狼优化决策树的大数据分类

本文方法大致可以分为 3 个步骤：首先将大数据输入 Map-Reduce 框架中，在 Map-Reduce 框架中进行简化；其次将简化后的大数据作为 SVM 分类器的输入，SVM 分类器会调整精简后的数据，对数据进行粗略分类；最后基于差异灰狼优化决策树，对 SVM 分类器输出的类标签进行更精确的分类.

2.1　Map-Reduce 框架

假设大数据的尺寸为 45×1 000(45 为属性，1 000 为记录)，将其输入到 Map-Reduce 中，在 Map-Reduce 框架中进行拆分、映射和缩小 3 个操作. 首先，根据框架中存在的 n 个映射器对给定数据进行拆分，为某些映射器提供 30×1 000，随后将其余 15×1 000 提供给其他映射器；其次在映射过程中，通过将接收的一组数据转换为另一组数据，同时将每个元素分别分解为键-值对. 在缩小过程中，它接受映射的结果作为关键字，并将这些数据元组合并为次要元组，此过程中采用主成分分析法[11]来减少数据量至 1×1 000 的维度.

2.2　SVM 分类

SVM 是一种基于统计学习理论的模式识别方法，该方法属于一种二分类模型，在解决小样本和高维非线性数据的模式识别问题中具有很大的优势. SVM 目的是在多维空间中找到一个能将全部样本单元分成两类的最优平面，这一超平面应使两类中距离最近的点的间距尽可能大. 两类正则可分问题的超平面构成一个 n 维特征空间，其数学定义可以表示为

$$P(\boldsymbol{A}) = \boldsymbol{E}^{\mathrm{T}}\boldsymbol{A} + s = 0 \tag{9}$$

式(9)中：$\boldsymbol{E}$ 表示法线向量，$\boldsymbol{A}$ 表示多维向量，s 表示从超平面到原点的距离.

当训练数据线性可分时，存在多个分离超平面将两类数据正确分开. 为了能够使解唯一，而且更具鲁棒性，要求间隔最大化. 但是当数据中的某些特征缺失时，向量数据不完整，会导致 SVM 泛化程度降低，训练结果变差，发生数据重叠现象，因此准确的训练数据分区是一项具有挑战性的功能. 为了应对这一情况，本文将 SVM 输出的分类结果进行优化处理.

2.3　基于差异灰狼优化的决策树

当大数据中遇到数据不平衡和数据类型缺失两种问题时，会导致模型误判，分类准确度降低. 为了解决这一问题，采用数据类别权重修改策略，将不重要数据的权重降低，将重要数据的权重升高. 权重 λ 的改变基于差异灰狼优化的最佳选择，并同时优化整数值 n.

决策树的构建通常分为特征选择、决策树生成、决策树修剪 3 个过程. 特征选择是构建决策树之前的重要步骤，如果是随机选择特征，那么所建立的决策树学习效率很低. 决策树生成始于一棵空树，要分类的样本从树根进入，在树的每个节点通过对样本某种属性的判断选择不同的路径逐步下降到底，得出其所属类别. 为了防止决策树变得过度拟合，在构造整棵树后尝试消除多余的节点，这个过程就是剪枝. 剪枝的过程就是对具有相同父节点的节点进行检查，判断将其合并后信息增益是否会小于某个指定发值. 若是，则合并这些节点.

2.3.1　特征选择

通常特征选择的准则是基于信息增益理论进行的，该理论为每个决策节点候选属性列表之间的选择提供了标准. 通过计算每个特征的信息增益，选取信息增益最大的特征

$$Gain(S, R_p) = Q(S) - Q_{R_p}(S) \tag{10}$$

式(10)中：R_p 表示样本特征，S 表示训练集.

$$Q(S) = -\sum_{j=1}^{k} \frac{k(B_j, S)}{|S|} \log_2 \frac{k(B_j, S)}{|S|} \tag{11}$$

$$Q_{R_p}(S) = \sum_{u \in C(R_p)} \frac{|S_u^{R_p}|}{|S|} Q(S_u^{R_p}) \tag{12}$$

式(11)、式(12)中：$k(B_j, S)$表示训练集S中属于B_j类的对象数，$C(R_p)$表示特征R_p的有限域，$|S_{u^p}^{R_p}|$表示特征R_p的值为u的对象集的基数. 信息增益比是指信息增益与训练数据集S关于特征R_p的值的熵之比，可以表示为

$$Gro(S, R_p) = \frac{Gain(S, R_p)}{SplitInfo(S, R_p)} \tag{13}$$

式(13)中：$SplitInfo(S, R_p)$表示将Q划分为n个子集所产生的潜在信息，数学定义为

$$SplitInfo(S, R_p) = \sum_{u \in C(R_p)} \frac{|S_{u^p}^{R_p}|}{|S|} \log_2 \frac{|S_{u^p}^{R_p}|}{|S|} \tag{14}$$

2.3.2 决策树生成

决策树的生成从根节点开始，通过计算节点处所有可能特征的信息增益，选取信息增益最大的特征作为节点特征，根据该特征的不同取值建立子节点. 对于子节点，采用上述方法进行递归. 当所有特征的信息增益均很小或没有特征可以选择时，决策树构建完成.

2.3.3 决策树剪枝

决策树的剪枝是为了防止过拟合现象出现，对已生成的决策树通过优化损失函数来去掉一些不必要的分类特征，降低模型的整体复杂度. 剪枝的方式是从树的叶节点出发，向上回缩，逐步判断. 如果剪掉某一特征后，整棵决策树所对应的损失函数更小，那就将该特征及带有的分支剪掉.

决策树的剪枝一般通过决策树整体损失函数的最小化来实现，损失函数可以表示为

$$C_a(T) = C(T) + a|T| \tag{15}$$

式(15)中：T表示任意子树，$|T|$表示子树的节点个数，$C(T)$表示训练数据的预测误差，a为参数，衡量模型的复杂度与训练数据的拟合程度.

2.3.4 优化方案和目标函数

本文的主要目的是构建决策树过程中优化决策树分类器，提出了一种基于差异灰狼优化的决策树优化策略，以决策树分类器的准确度为目标函数，通过对数据类别权重的优化修改，构建最佳决策树. 优化方案的目标函数为

$$OB = \max(\text{accuracy}) \tag{16}$$

式(16)中：accuracy表示决策树分类器的准确度.

目前，GWO算法已经解决了许多工程问题，但是GWO算法也存在一定的局限性，缺乏寻找影响算法收敛速度的全局最优解的能力. 模型复杂度高，其本质上是非线性的，由于收敛因子线性降低，无法真正反映实际的搜索过程. 为了克服常规GWO算法在效率、探索和开发特性方面的局限，本文开发了差异灰狼优化算法. 在常规GWO算法的基础上，潜在解位置的更新方程(7)修改为

$$\vec{Z}(x+1) = \left[\frac{\vec{Z}_1 + \vec{Z}_2 + \vec{Z}_3}{3}\right] - \left[\frac{\vec{Z}_1' + \vec{Z}_2' + \vec{Z}_3'}{3}\right] \tag{17}$$

式(17)中：$\vec{Z}_1'$，$\vec{Z}_2'$和$\vec{Z}_3'$分别表示第一，第二和第三最差解.

3 评估指标与实验结果分析

为了验证本文分类算法的可行性及有效性，利用大量数据在JAVA中进行实验，并且将本文模型的性能与灰狼优化决策树算法(GWO)、蚁群优化-人工神经网络联合算法(Ant Colony Optimization and Artificial Neural Network algorithm, ACO-ANN)[12]、无尺度二元粒子群算法(Scale-Free Binary Particle Swarm Optimization, SFBPSO)[13]和模糊最小最大值神经网络与脑风暴优化联合算法(Fuzzy Min-Max neural network and Brain Storm Optimization, FMM-BSO)[14]等进行了比较.

3.1 评估指标

对一个二分类问题，实验时只存在正、负两种取值. 当一个样本为正类，且被预测为正类时，称为真正

类(TP)；假若负类被预测为正类，称为假正类(FP)；假若是负类被预测成负类，称为真负类(TN)；假若正类被预测为负类，称为假负类(FN)．本文采用准确度($Accuracy$)、真正例率(TPR)、假正例率(FPR)和 F_1 分数(F_1-$score$)4 个指标评估模型的性能，4 个指标的数学定义由式(18)—式(21)给出．

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{18}$$

$$TPR = \frac{TP}{TP + FN} \tag{19}$$

$$FPR = \frac{FP}{FP + TN} \tag{20}$$

$$F_1 - score = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{21}$$

3.2　实验结果分析

为了有效验证算法的性能，实验采用两种方式：①对于 Cleveland 和 Hungarian 两个数据集，即从公开来源的加州大学欧文分校(University of California Irvine，UCI)机器学习存储库网页中收集的心脏病数据集[15]，通过改变组合数据集中训练数据的百分比来对本文方法进行分析；②利用组合数据集对模型进行测试．实验结果发现，差异灰狼优化算法比标准 GWO 具有一定的优势，4 种测试指标均有明显提高．同时，本文的算法比其他分类算法具有更好的效果．

表 1 给出了利用组合数据集进行测试的结果对比，从表 1 中可以看出，本文的分类方法整体性能优越，相对于其他分类算法在性能方面均有明显的提升．

表 1　组合数据集的测试结果对比

指标	SFBPSO	ACO-ANN	FMM-BSO	GWO	本文方法
准确度	0.586	0.667	0.698	0.623	0.713
TPR	0.628	0.679	0.711	0.685	0.802
FPR	0.383	0.342	0.295	0.346	0.207
F_1 分数	0.641	0.692	0.727	0.663	0.753

4　结　　语

本文提出了一种基于差异灰狼优化决策树的大数据分类方法，用于解决现有大数据分类算法中存在准确率低和复杂度高的问题．针对医疗大数据中存在的大量小样本数据和非结构化数据，本文方法首先在 Map-Reduce 框架中对复杂的大数据进行降维简化处理，然后利用支持向量机对简化后的数据进行粗略分类．为了获得更准确的分类效果，最后采用基于差异灰狼优化的决策树对 SVM 输出的类标签作精细化处理．实验结果表明，本文方法在复杂大数据方面优于其他分类算法．

参考文献：

[1] ELHOSENY M, ABDELAZIZ A, SALAMA A S, et al. A Hybrid Model of Internet of Things and Cloud Computing to Manage Big Data in Health Services Applications [J]. Future Generation Computer Systems, 2018, 86: 1383-1394.

[2] WANG Y C, KUNG L, WANG W Y C, et al. An Integrated Big Data Analytics-Enabled Transformation Model: Application to Health Care [J]. Information & Management, 2018, 55(1): 64-79.

[3] TAN J Y, XIONG T, MIAO H X, et al. A Case Study of Medical Big Data Processing: Data Mining for the Hyperuricemia [C]//2018 IEEE 3rd International Conference on Cloud Computing and Big Data Analysis (ICCCBDA). Chengdu: IEEE, 2018.

[4] NAIR L R, SHETTY S D, SHETTY S D. Applying Spark Based Machine Learning Model on Streaming Big Data for Health Status Prediction [J]. Computers & Electrical Engineering, 2018, 65: 393-399.

[5] HU Y, DUAN K, ZHANG Y, et al. Simultaneously Aided Diagnosis Model for Outpatient Departments via Healthcare

Big Data Analytics [J]. Multimedia Tools and Applications, 2018, 77(3): 3729-3743.

[6] VENTURA-MOLINA E, ALARCÓN-PAREDES A, ALDAPE-PÉREZ M, et al. Gene Selection for Enhanced Classification on Microarray Data Using a Weighted K-NN Based Algorithm [J]. Intelligent Data Analysis, 2019, 23(1): 241-253.

[7] KHAN S, KHAN A, MAQSOOD M, et al. Optimized Gabor Feature Extraction for Mass Classification Using Cuckoo Search for Big Data E-Healthcare [J]. Journal of Grid Computing, 2019, 17(2): 239-254.

[8] DAS H, NAIK B, BEHERA H S. Medical Disease Analysis Using Neuro-Fuzzy with Feature Extraction Model for Classification [J]. Informatics in Medicine Unlocked, 2020, 18: 100288.

[9] NAGARAJAN G, DHINESH BABU L D. An Empirical Comparison of Six Supervised Machine Learning Techniques on Spark Platform for Health Big Data [M]//Smart Intelligent Computing and Applications. Singapore: Springer, 2019.

[10] 周孟然, 卞 凯, 刘卫勇, 等. 属性约简结合GWO-SVC的乳腺恶性肿瘤数据诊断研究 [J]. 计算机应用与软件, 2019, 36(8): 155-159, 234.

[11] 妥 娅, 武建辉, 赵永成. 主成分分析法和BP神经网络组合模型在天津市某医院住院费用研究中的应用 [J]. 医学与社会, 2018, 31(2): 45-47, 54.

[12] JOSEPH MANOJ R, ANTO PRAVEENA M D, VIJAYAKUMAR K. An ACO-ANN Based Feature Selection Algorithm for Big Data [J]. Cluster Computing, 2019, 22(2): 3953-3960.

[13] GUPTA S L, BAGHEL A S, IQBAL A. Big Data Classification Using Scale-Free Binary Particle Swarm Optimization [M]// Harmony Search and Nature Inspired Optimization Algorithms. Singapore: Springer, 2019.

[14] POURPANAH F, LIM C P, WANG X Z, et al. A Hybrid Model of Fuzzy Min-Max and Brain Storm Optimization for Feature Selection and Data Classification [J]. Neurocomputing, 2019, 333: 440-451.

[15] PADMANABHAN M, YUAN P Y, CHADA G, et al. Physician-Friendly Machine Learning: a Case Study with Cardiovascular Disease Risk Prediction [J]. Journal of Clinical Medicine, 2019, 8(7): 1050.

Big Data Classification Method Based on Optimized Decision Tree by Differential Grey Wolf Optimization

LYU Ting-qin, WEI Meng

School of Information Science and Technology, Zhengzhou Normal University, Zhengzhou 450044, China

Abstract: Aiming at the problems of low accuracy in the existing big data classification algorithms, a big data classification method based on optimized decision tree by differential grey wolf optimization has beenproposed. This methodis first used to inputcomplex big data into the Map-Reduce framework, andto reducethe input data using principal component analysis. Then itis used tosupport vector machines to classify the compressed dataroughly. Finally, it uses a decision based on the difference gray wolf optimization. The tree finely classifies the class labels output by the support vector machine to obtain higher classification accuracy. Experimental results show that compared with other classification algorithms, the proposed method has obvious advantages in the classification of complex big data.

Key words: big data; data classification; differential grey wolf optimization; decision tree

责任编辑 夏 娟