基于 FDR 的证据理论改进算法

丁烈骁

Big Data Analytics Trading Inc. 美国

【摘要】为简化证据理论合成规则融合过程，提高其融合效果，本文应用特征降维（Feature Dimension Reduction，FDB）技术，提出一种行之有效的证据理论改进算法。实验结果表明：基于 FDR 的证据理论改进算法具有融合过程简单、融合效果好、类型识别率高等特点，该算法经过数据集测试后，其类型识别率升高至 94%，完全符合实际应用需求。希望通过这次研究，为相关人员提供有效的借鉴和参考。

【关键词】证据理论；组合规则；样本分类

【收稿日期】2023 年 8 月 6 日 【出刊日期】2023 年 9 月 27 日 【DOI】10.12208/j.aics.20230030

Improved algorithm of evidence theory based on FDR

Liexiao Ding

Big Data Analytics Trading Inc., United States of America

【**Abstract**】In order to simplify the fusion process of evidence theory synthesis rules and improve its fusion effect, this paper applies Feature Dimension Reduction (FDB) technology to propose an effective evidence theory improvement algorithm. The experimental results show that the improved algorithm based on FDR evidence theory has the characteristics of simple fusion process, good fusion effect, and high type recognition rate. After being tested on the dataset, the type recognition rate of the algorithm increased to 94%, fully meeting the practical application requirements. I hope to provide effective reference and guidance for relevant personnel through this study.

【**Keywords**】Evidence theory; Combination rules; Sample classification

证据理论作为一种有效工具，主要用到模糊决策思想，在目标识别、决策制定等领域具有重要作用。在进行特征融合之前，为促使样本特征融合过程变得更加简单化，需要筛选、降维处理原始样本特征。另外，为避免因样本特征权重分配不合理而降低最终实验结果精确度，需要在进行特征融合期间，将较低权重分配给存在矛盾关系的样本特征[1]。本文提出一种基于 FDR 的证据理论改进算法，该算法在正式融合样本特征之前，筛选和删除对分类结果影响程度较低的特征，不断降低初始样本特征数量，促使融合过程变得更加简单化，从而获得最佳融合效果[2]。所以，强化对基于 FDR 的证据理论改进算法研究显得尤为重要。

1 基础理论

1.1 D-S 证据理论

D-S 证据理论主要是由 Dempster 和 Shafer 两名学者所提出的证据理论，运用该理论，可以从模糊、不确定的信息中，制定出相关决策。该证据理论被广泛地应用于各个领域中，并取得了良好的应用效果[3]。D-S 证据理论在具体应用时，使用 2^n 代表整个空间中的各个子集集合。子集集合表示如下：

$$2^n = \{\phi, \{\theta_1\}, \{\theta_2\}, \cdots \{\theta_n\}, \{\theta_1, \theta_2\}, \cdots \{\theta_1, \theta_2, \cdots \theta_n\}\} \quad (1)$$

在 D-S 证据理论中，通常会用到质量函数，该函数符合在 $2^n \to [0,1]$ 上的映射。运用上述公式，可以保证证据融合效果，简化证据融合过程，但是，当某一证据支持度过低或者达到 0 时，其最终融合效果较差，不够理想。

为解决以上问题，技术人员要引入和应用证据权值，不断降低此类证据的影响程度，此外，还要在整个融合样特征之前，筛选和删除支持度较低的样本特征。

作者简介：丁烈骁（1987-）男，硕士，计算机网络架构工程师，研究方向：计算机网络架构

1.2 特征选择

在各个样本特征之间，通常会存在较高的关联度，这就增加计算消耗量。此外，部分特征一旦处理不当，会降低最终预测结果精确度[4]。所以，在选择样本特征时，需要从各个样本特征中，选择出含有部分特征的样本，并将其设置为新样本特征，在选择整个过程中，无论是否改变特征值，选择后的特征纬数必然有所减少。目前，主要用到以下两种常用的特征选择方法[5]。

（1）均方差分析法

俊芳差分析法在具体运用时，需要设置某一特定的方差阈值，并筛选和删除小于阈值的特征值。如果方差阈值没有被指定，需要对各个非零方差特征进行全部保留，并筛选和删除含有相同值的样本特征值]。总之，当样本数据集特征值较小时，优先选用均方差分析法。借助方差值，分析样本数据集特征时，要将其与极差、标准差进行分析和对比，方差最终计算结果改变，会直接影响样本特征波动性，某一样本特征值改变，直接影响样本类型划分结果。借助方差分析样本数据集时，需要对数据集全部数据进行统一化分析和处理，当数据量和样本特征量不断增加时，整个计算过程会变得越来越繁琐和复杂。所以，均方差分析法在具体应用时，存在一定的局限性，仅仅适用于样本特征量较小的数据集领域中。

（2）主成分分析法

主成分分析法作为一种数据集常用方法，具有简单高效特点，运用该分析方法，不仅可以将数据集纬数降到最低，还能简化原始数据结构，有效地降低数据损失程度。此外，运用主成分分析法，可以筛选和删除回归分析中所产生的特征数量。对于高纬数据集而言，大量特征之间存在一定的线性关系，运用该分析法，可以将相关特征数量和样本特征件关联度降到最低。采用主成分分析法，对高温数据集进行分析时，首先，要计算出多个主成分贡献率之和，确保贡献率达到较高状态。其次，降维处理后的样本特征存在含义模糊特点，当样本特征数量过小时，不宜采用主成分分析法，当数据集处于高纬状态时，可以选用主成分分析法。

2 基于 FDR 的证据理论改进算法

当原始样本数据集特征降维处理后，相关特征数量明显降低，此时降维数据形成新的 BPA。然后，采用方差分析法、对简单数据进行处理，从而筛选出方差超过阈值的数据集特征。运用主成分分析法，对复杂数据集进行处理，并对选择好的样本特征进行降维处理。整个算法过程中，将原始数据集矩阵设置为输入值；将降维数据集矩阵设置为输出值。改进算法流程如图 1 所示，特征选择和降维处理是整个改进算法的重要环节。

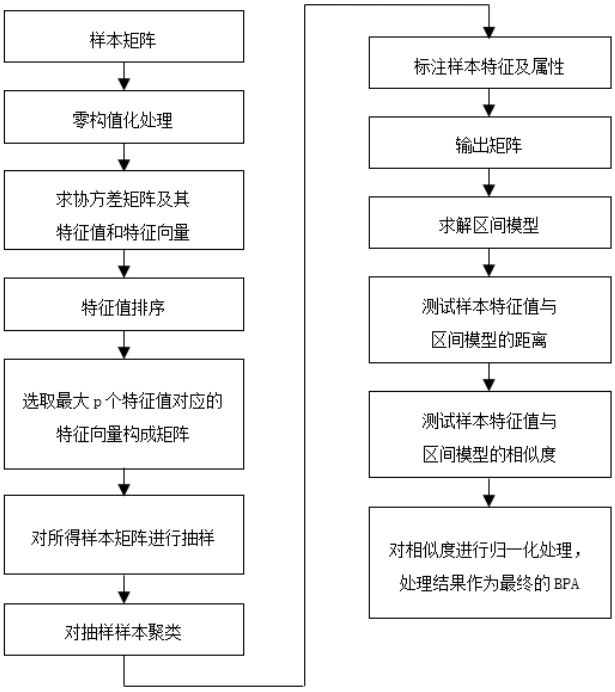

图 1 算法流程

3 实验

3.1 数据集描述

在本次实验期间，将 20 万名 Instacat 用户所产生的 400 多万份杂货订单样本设置为数据集，为避免数据集被恶意入侵和篡改，需要对其进行匿名处理，这些数据集可以详细地描述客户订单与时间两者之间的影响关系。所有用户需要按照指定的顺序，购买相应订单中的产品。同时，这些数据集还为用户提供订单下达时间以及相关时间度量。

3.2 数据库介绍

客户、产品、订单等各个实体均含有唯一 ID 标识。数据表与变量名称真实存在。数据表如表 1 所示，从表 1 中可以看出，Products.csv 数据表表示商品信息；Order_products_prior.csv 数据表内容主要是由订单信息和商品信息组成；Aisles.csv 数据表主要描述商品所属具体物品类别。

3.3 原始数据降维处理

在进行原始数据降维处理时，首先，从以上数据库中读出 4 张常用数据表，借助表中的 ID 标识，合并处理多张数据表，然后，选择 500 组合并数据，并对其进行实验。借助主成分分析算法，降维处理数据集特征，降维处理后的数据集主要包含原始数据集的 90%以上的信息。特征数量出现大幅度下降，下降至 27 个，促使原始数据集特征融合过程变得越来越简单。利用 K 均值聚类（K-means）算法对降维处理后的数据集进行聚类分析，并从这些数据集中选取前 500 个样本，将其划分为五类用户，并用数字 0、1、2、3、4 对其进行表示，所获得的样本聚类结果和聚类分布。最后，进一步分析和对比上述聚类结果，并对聚类数据集所对应的轮廓系数进行计算，其计算公式如下：

$$sc_i = \frac{b_i - a_i}{\max(b_i, a_i)} \tag{2}$$

（2）式中的 i 代表已知聚类数据中的样本；b_i 代表 i 与其他簇各个样本所对应的平均距离；a_i 代表 i 与本身簇之间所对应的距离平均值。经过计算，发现其轮廓系数为 0.6127，位于-1~1 之间，当轮廓系数无限接近于 1 说明数据集的内聚度和分离度达到在最佳状态。

3.4 对降维后的数据集生产 BPA 进行分类识别

当数据集降维处理后，需要采用分类识别方法，对所形成的 BPA 进行识别处理。首先借助降维处理后的数据集，对数据库进行验证，然后，从五种用户类型中，选出 30 个样本，每种用户类型各占 6 个，从整个样本中选出最大值和最小值，完成对区间数模型构建。区间数据模型如表 2 所示。其次，通过对测试样本与区间数模型之间的相似度进行计算，求解出六种特征值所对应的 BPA，然后，运用证据理论组合规则，对其特征进行融合处理，从而获得最终特征融合结果。最后，系统化分析融合结果，发现融合结果可以直接决定和影响测试样本类型，BPA 值最大的类型就是待测样本所对应的类别。

表 1 数据说明

数据表名称	Products.csv	Order_products_prior.csv	Orders.csv	Aisles.csv
表内容说明	商品信息	订单与商品信息	用户的订单信息	商品所属具体物品类别

表 2 区间数据模型

类型	特征 1	特征 2	特征 3	特征 4	特征 5	特征 6
类型 0	[-24.3,6.4]	[-16.9,9.6]	[-14.3,8.0]	[-19.8,10.9]	[-28.3,14.4]	[-22.7,8.4]
类型 1	[-5.0,7.9]	[-11.3,14.1]	[-13.9,12.0]	[-12.1,10.9]	[-12.7,15.6]	[-25.8,9.2]
类型 2	[-12.0,7.4]	[29.5,15.8]	[-28.2,10.7]	[-9.50,17.0]	[-7.22,12.7]	[-20.4,8.9]
类型 3	[-19.3,25.4]	[-4.50,3.0]	[-16.4,9.5]	[-9.2,17.8]	[-27.3,19.8]	[-24.1,9.9]
类型 4	[14.3,9.4]	[-9.3,10.2]	[-7.5,11.1]	[-11.5,16.4]	[-16.1,18.6]	[-19.0,9.7]

3.5 实验结果分析

在本次实验中，为有效地检验和分析本文改进算法在数据集中的分类识别效果，将选取的支持度系数设置为 3，结合如表 2 所示的区间数模型，测试

30 个待测样本数据集，发现其平均类型识别率达到 94%。

4 结束语

综上所述，本文提出的基于 FDR 的证据理论改进算法，主要运用 Instacart 数据集进行算法检验，与其他传统算法相比，本文改进算法的提出和应用，可以促使证据融合过程变得更加简单化，同时，还降低原始样本中特征之间关联度，保证最终融合效果。实验结果表明：运用本文改进算法，可以实现对分类问题的精确化识别，识别率高达 94%。但是，本次研究工作还存在以下需要完善的地方：如运用特征降维方法，改进权重分配方案，只有这样，才能获得最佳证据融合效果。

参考文献

[1] 徐吉辉,史佳辉,陈玉金,等.基于云模型和证据理论的现役装备改进方案评价[J].火力与指挥控制,2022,47(2):25-31.

[2] 高天昊,曲卫,董尧尧,等.基于改进 D-S 证据理论的 MPAR 行为状态识别方法[J].电光与控制,2022,29(12):1-6.

[3] 任荣明,谭海立,罗月静.基于改进证据理论和检测信息的混凝土桥梁服役状态评估分析[J].西部交通科技,2021(10):120-123,181.

[4] 田文杰,徐吉辉,祝娜,等.基于 Z-number 和改进 DS 证据理论的风险评估方法[J].火力与指挥控制,2023,48(1):43-49.

[5] 游昊,石恒初,杨远航,等.基于改进 D-S 证据理论的电网故障多源信息智能融合诊断方法[J].广东电力,2020,33(11):16-25.