

网络大数据下的冗余数据分类优化算法研究

欧阳柏成

(贵阳学院 数学与信息科学学院, 贵州 贵阳 550005)

摘 要: 提出了基于模糊支持向量机算法网络大数据下的冗余数据分类优化方法。提取网络大数据环境下的冗余数据属性特征, 为冗余数据分类提供准确的数据基础。根据模糊支持向量机相关理论, 获取最优分类平面, 从而实现冗余数据分类优化处理。实验结果表明, 利用改进算法进行网络大数据下的冗余数据分类优化处理, 能够提高分类的准确性, 取得了令人满意的效果。

关键词: 网络; 大数据; 冗余数据

中图分类号: TP311

文献标识码: A

文章编号: 1000-7180(2015)01-0128-03

Research on the Redundancy Data Classification Optimization Algorithm under the Big Data of Network

OUYANG Bo-cheng

(College of Mathematics and Information Science, Guiyang University, Guiyang 550005, China)

Abstract: The fuzzy support vector machine (SVM) algorithm is proposed under the large data of network redundancy data classification optimization method. Extract the network under the environment of large data redundancy data attributes, providing accurate data basis for redundant data classification. According to the related theory, fuzzy support vector machine (SVM) to obtain the optimal classification plane, so as to realize optimization of redundant data classification. The experimental results show that the improved algorithm under the large data of network optimization of redundant data classification, can improve the accuracy of classification, satisfactory results have been achieved.

Key words: network; Big data; Redundant data

1 引言

随着计算机网络规模的迅速扩大, 网络中数据的规模也越来越庞大^[1]. 利用数据分类算法进行网络数据分类已经成为一种重要的数据获取方法^[2]. 数据分类方法是计算机网络数据提取的核心部分^[3]. 网络数据中存在大量的冗余数据, 利用数据分类方法, 能够将大量的冗余数据进行合理分类, 从而实现有效网络数据的准确提取, 提高网络运行效率^[4]. 这种数据分类方法在数据库领域、人工智能领域等相关领域都有极高的应用价值, 受到很多专家的关注^[5]. 因此, 数据分类算法已经成为网络数据领

域需要研究的热点课题^[6]. 当前阶段, 主要的网络冗余数据分类算法主要有基于支持向量机的网络冗余数据挖掘方法、基于小波变换算法的网络冗余数据分类方法和基于数据增益算法的网络冗余数据分类方法^[7]. 其中, 最常使用的是基于小波变换算法的网络冗余数据分类方法^[8]. 由于数据分类方法的应用范围非常广泛, 因此成为很多专家研究的重点课题, 拥有广阔的发展空间^[9].

在网络冗余数据分类的过程中, 由于网络数据量非常庞大, 其中包括大量的冗余数据, 将会造成冗余数据之间的关联性下降, 导致网络的运行效率降低.

收稿日期: 2014-03-27; 修回日期: 2014-05-21

基金项目: 贵州省科技厅联合基金项目(黔科合 LH 字[2014]7197 号)

为了避免上述传统方法的缺陷,提出基于模糊支持向量机算法网络大数据下的冗余数据分类优化方法.提取网络大数据环境下的冗余数据属性特征,为冗余数据分类提供准确的数据基础.根据模糊支持向量机有关理论,获取最优分类平面,从而实现冗余数据分类优化处理.

2 网络冗余数据分类优化方法的提出

利用传统算法进行网络大数据下冗余数据分类的过程中,无法避免由于大量的冗余数据造成冗余数据之间的关联性下降,造成致网络的运行效率降低的缺陷,为此,提出了基于模糊支持向量机算法网络大数据下的冗余数据分类优化方法.

2.1 冗余数据属性特征提取

利用局部特征分析方法,能够提取冗余数据的关键特征,对其进行分析,将其作为冗余数据分类的数据基础.这种特征方法不受数据采集的过程中网络延迟变化的影响,通过冗余数据关键特征与邻近区域的数据特征值进行比较,能够描述冗余数据的特征.

设置冗余数据的相性度能够用 $E_S^{\min}$ 进行描述,则可以得到下述公式:

$$E_S^{\min}(M) = \min\{E_S(M, M^{\text{uniform}}), \text{all } M^{\text{uniform}}\}$$

$$E_S(M, M^{\text{uniform}}) = \min\{\varphi(S(M_i), M^{\text{uniform}})\} \quad (1)$$

2.2 网络海量数据分类过程

根据最优分类平面的相关原理能够对冗余数据进行分类处理.将冗余数据分类问题转化为最有平面求解的问题,利用下述公式能够描述需要求解的问题:

$$R(\beta) = \sum_{j=1}^p \beta_j - \frac{1}{2} \sum_{j,k=1}^p \beta_j \beta_k z_j z_k (y_j \cdot y_k) \quad (2)$$

上述最优分类平面求解问题的需要满足下述约束条件:

$$\begin{cases} \sum_{j=1}^p z_j \beta_j = 0 \\ \beta_j \geq 0, j = 1, 2, \dots, p \end{cases} \quad (3)$$

假设冗余数据中的特征出现非线性变换,则需要利用内积 $L(y, y^T)$ 替代最优分类函数中的点积.则最优分类平面求解问题能够变换后的结果能够用下述公式进行描述:

$$R(\beta) = \sum_{j=1}^p \beta_j - \frac{1}{2} \sum_{j,k=1}^p \beta_j \beta_k z_j z_k L(y_j \cdot y_k) \quad (4)$$

利用下述公式能够描述上述公式对应的最优分类函数:

$$g(y) = \text{sgn}\left(\sum_{j=1}^p \beta_j z_j L(y_j \cdot y_k) + c'\right) \quad (5)$$

最优分类平面方法能够对两种不同的种类进行分类,而冗余数据分类问题是一个多种类分类问题.因此,需要将网络中大数据下冗余数据分类问题转换为多个二分类的问题.并分别对上述问题进行求解.从而获取网络大数据下的冗余数据分类结果.当前阶段,主要有两种分类问题的转换方法,第一种方法是一对多的分类方法,第二种方法是一对一的分类方法.由于网络大数据下的冗余数据的运算复杂程度较高,冗余数据的特征数量也较多,因此,需要采用一对一的分类方法进行网络下大数据下的冗余数据分类问题的转换处理.

3 实验结果及分析

为了验证改进算法的有效性,需要进行一次实验.利用仿真软件 matlab7.1 构建实验环境.设置网络数据库中所有数据的数目是 1 000 个.冗余数据特征种类数目是 20 种.

在冗余数据种类较少的情况下,利用不同算法进行 10 次冗余数据分类,将实验过程中的数据进行整理分析,能够得到表 1 中的实验结果.

表 1 冗余数据种类较少时不同算法数据表

试验次数	支持向量机 算法分类准 确性/%	小波变换算 法分类准确 性/%	数据增益算 法分类准确 性/%	模糊支持向 量基算法分 类准确性/%
1	96	95	94	97
2	97	92	92	98
3	95	91	89	97
4	94	94	91	96
5	94	93	93	94
6	93	89	90	95
7	95	91	91	93
8	94	88	87	94
9	94	93	89	96
10	95	92	93	95

在冗余数据种类较多的情况下,利用不同算法进行 10 次冗余数据分类,将实验过程中的数据进行整理分析,能够得到表 2 中的实验结果.

通过上述实验能够得知,利用改进算法进行网络大数据下冗余数据分类,能够避免由于数据库中的冗余数据之间相关性较差的缺陷,从而提高了冗

余数据分类的效率.

表 2 冗余数据种类较多时不同算法数据表

试验 次数	支持向量机 算法分类准 确性/%	小波变换算 法分类准确 性/%	数据增益算 法分类准确 性/%	模糊支持向 量基算法分 类准确性/%
1	89	78	69	98
2	87	72	67	98
3	92	78	67	98
4	85	77	65	96
5	83	73	62	97
6	87	75	65	96
7	85	72	65	96
8	89	69	60	94
9	82	65	50	94
10	78	58	42	92

4 结束语

针对传统算法在进行网络大数据下冗余数据分类过程中,无法避免由于网络数据量较大造成数据之间的相关性降低,导致网络运行效率降低的弊端.为此,提出了基于模糊支持向量机算法网络大数据下的冗余数据分类优化方法.提取网络大数据环境下的冗余数据属性特征,为冗余数据分类提供准确的数据基础.根据模糊支持向量机相关理论,获取最优分类平面,从而实现冗余数据分类优化处理.实验结果表明,利用改进算法进行网络大数据下的冗余

数据分类优化处理,能够提高分类的准确性,取得了令人满意的效果.

参考文献:

- [1] 毕长剑,大数据时代建模与仿真面临的挑战[J]. 计算机仿真,2014,31(1): 1-3.
- [2] 万常选,刘喜平,著. XML 数据库技术. 北京:清华大学出版社. 2008;137-200.
- [3] 孙一中. XML 理论和应用基础[M]. 北京:北京邮电大学出版社. 2000;130-134.
- [4] 韩彤. 跟踪和管理并发用户数量提高数据库系统效率[J]. 信息技术与信息化,2011(5):49-50.
- [5] 隆茜,郭劲赤. 985 高校图书馆自建数据库的调查与分析[J]. 图书馆学研究,2010(18):27-31.
- [6] 韩彤. 跟踪和管理并发用户数量提高数据库系统效率[J]. 信息技术与信息化,2011(5):49-50.
- [7] 江梅. SciFinder Scholar 数据库与 CA on CD 数据库比较[J]. 现代情报,2006(3):83-86.
- [8] 宋淑彩. 面向 Web 的数据挖掘技术在网站优化中的个性化推荐方法的研究与应用[J]. 科技通报,2012,2(28):118-119.
- [9] 邓乃扬,田英杰. 数据挖掘中的新方法—支持向量机[M]. 北京:科学出版社,2004(6):36-39.

作者简介:

欧阳柏成 男,(1976-),硕士研究生,副教授.研究方向为数据库技术、网络工程.

(上接第 127 页)

- [5] Gao Ya, Qian Zheng, Cheng Kai. An implementation and verification of wireless VoIP communication based on WINCE 6.0 applied in underground mine[C]//IM-CCC. Hangzhou, China; IEEE, 2012:1637-1640
- [6] 蒋峰,张凌涛,贺超英. WiFi 技术在矿井远程监控系统中的应用[J]. 煤矿安全,2010(3):62-65.
- [7] 陈许建,白创,彭浩,等. 2.45GHz RFID 标签设计与实现关键技术研究[J]. 微电子学与计算机,2013,31(8):147-150.
- [8] 李苏阳. 光纤工业以太环网在煤矿的应用[J]. 煤炭科技,2012(4):62-63.
- [9] Ekici Eylem, Vuran Mehmet Can. Wireless heterogeneous networks and next generation internet[J]. Mobile Networks and Applications, 2010;607-609
- [10] 马飞,杨飞霞,徐光宪. 网络与终端协同选择及切换机制的研究[J]. 计算机科学,2012,39(1):77-81.

- [11] 高旭东. 物联网(智能家居)中无线异构网络融合网关的设计与应用[D]. 南京:南京邮电大学,2013.
- [12] 吴作顺. WSN 与计算机网的融合模型研究[J]. 计算机科学,2010(2):94-96.
- [13] Samsung Electronics Coltd. S3C2440A 32-BIT CMOS Microcontroller Users Manual[R]. Revision 1. Korea, 2004.
- [14] 杨顺,章毅,陶康. 基于 ZigBee 和以太网的无线网关设计[J]. 计算机系统应用,2010,19(1):194-197.

作者简介:

石玲玲 女,(1990-),硕士研究生.研究方向为计算机控制. E-mail:15855401823@163.com.
李敬兆 男,(1964-),教授,博士生导师.研究方向为嵌入式技术、物联网和云计算.