

数据挖掘中特征选择算法研究

荣盘祥, 曾凡永, 黄金杰

(哈尔滨理工大学 自动化学院, 黑龙江 哈尔滨 150080)

摘 要: 针对在数据挖掘过程中存在的数据冗余特征和维灾难问题, 依据 ReliefF 算法和主成分分析算法的理论基础方法, 建立了基于 ReliefF 优化的核主成分分析的二次特征选择法, 并给出了该方法的实验结果. 该方法能够有效处理维度过高、具有冗余和无关特征的数据, 结合机器学习算法, 使数据挖掘系统得到准确高效的执行结果, 为决策人员提供有力的决策依据. 通过实验得出该算法具有更高的分类准确度的结论.

关键词: 数据挖掘; 特征选择; 主成分分析

DOI: 10.15938/j.jhust.2016.01.021

中图分类号: TP301.6 **文献标志码:** A **文章编号:** 1007-2683(2016)01-0106-04

Research On Feature Selection Algorithm In Data Mining

RONG Pan-xiang, ZENG Fan-yong, HUANG Jin-jie

(School of Automation, Harbin University of Science and Technology, Harbin 150080, China)

Abstract: Aiming at data redundancy and curse of dimensionality in data mining process, in accordance with the theoretical bases and methods of ReliefF algorithm and principal component analysis algorithm, this paper establishes the quadratic feature selection method on the basis of ReliefF optimization and principal component analysis, and gives out the experimental results of this method. This method can effectively process the data with high dimension, redundant and irrelevant features. Combined with machine learning algorithm, it makes the data mining system get an accurate and efficient implementation result, thus providing a solid decision-making foundation for decision makers. The conclusion is that this algorithm has a higher classification accuracy obtained through experiment.

Keywords: data mining; feature selection; principal component analysis

0 引 言

在 60 年代, Lewis 是最早开始讨论特征选择^[1-2]问题的研究人员, 当时的研究的重点是统计

学和信号处理. 通常只能获得相对较少的特征数目, 而且假定特征之间相互没有关系. 1992 年, Kira 和 Rendell 提出 Relief^[3-4] 特征选择算法及改进版本, 选取权重比较大的特征, 其余特征将被提出. 从 1990 之后, 我国的陈彬等人就提出并证明了特征选

收稿日期: 2015-06-10

基金项目: 黑龙江省博士后资助项目(LBH-Q11081); 黑龙江省教育厅科学技术研究项目(11551093).

作者简介: 荣盘祥(1965—), 男, 教授, 硕士研究生导师, E-mail: rongpx@hrbust.edu.cn;

曾凡永(1987—), 男, 硕士研究生;

黄金杰(1967—), 男, 教授, 博士研究生导师.

择算法中的最优特征子集^[8]是 NP-hard 问题,这对于特征选择问题的研究具有重要意义。

近些年来,对社交网络环境的数据挖掘^[6-8]越来越重视,经过数据的预处理怎样能得到较高的、有利数据发现的数据集,将是我们研究的焦点。本文依据 ReliefF^[9] 算法和主成分分析^[10] 算法的理论基础方法,建立了基于 ReliefF 优化的核主成分分析^[11] 的二次特征选择法并首先应用于社交网络数据挖掘。

1 基于对 ReliefF 优化的核主成分分析的二次特征选择算法

因为 Relief 算法只能删除对区分类别能力不强的特征。特征权重^[12] 的变化受到过程重复次数的影响,但是变化的条件是特征和分类或者类别相关。基于这个原因,可以通过优化 ReliefF 算法来准确评价多特征影响分类性能的程度。一个评价标准可能难以实现,所以提出了基于 ReliefF 算法的二次特征选择法,由于同类样本的间距小,首先需要将数据集进行有效的分离,而不同类样本间的距离大,接着我们用 ReliefF 算法得到样本的每个特征所对应的权重,权重越高表示这个特征和类关系越密切。接着利用 KPCA 对各个特征加权,代表对分类的影响程度不同。这样具有强化重要的特征,而弱化不相关的特征。最后利用搜索算法中计算特征之间相关度^[13] 的函数运用到整个特征识别系统中,去除了冗余的特征。因此去除掉这些与分类无关的特征既减少了运算量也兼顾了不相关的特征和冗余特征^[14] 的处理。以下是算法的流程图 1。

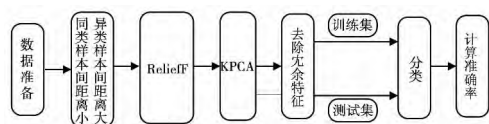


图 1 ReliefF 优化的核主成分分析的二次特征选择法算法流程图

1.1 ReliefF 算法

步骤一: 初始化特征的权重都为 0, FS 为空集。

步骤二: For $i = 1, i < m$; 在所有样本中随机选择某个样本 R ; 在 R 属于同类的样本中, 找到和 R 最近的相邻的样本 H_j . 在 R 不属于同类的样本中, 找到和 R 最近的相邻的 k 个样本 $M_j(C)$; 然后 For $A = 1, A < M$.

$$W(A) = W(A) - \frac{\sum_{j=1}^k \text{diff}(A, R, H_j)}{mk} + \frac{P(C)}{1 - P(\text{Class}(R))} \frac{\sum_{j=1}^k \text{diff}(A, R, M_j(C))}{mk} \quad (1)$$

步骤三: For $A = 0, A < N$; Do if $W(A) > \delta$ 将 A 个特征加到 FS 中。

ReliefF 算法中, $\text{diff}(A, R_1, R_2)$ 代表在特征 A 上, 实例样本 R_1 和 R_2 的差值, $M_j(C)$ 代表在分类 C 上的第 j 个最邻近的样本。

ReliefF 算法运算效率比较高, 而且所使用的数据没有类型限制, 但是对特征之间的关系不能进行主动分析, 对于冗余的特征不能去除。有可能某个特征是冗余的特征, 但是还是给这个冗余的特征赋予了权重。

1.2 基于核函数的主成分分析法

假设有 $\Phi(x_i)$, 即线性映射函数, 有 N 个样本点。将这输入的 N 的样本点转化到高维线性的特征空间 P 中, 假设如下:

$$\sum_{k=1}^N \Phi(x_k) = 0 \quad (2)$$

在高维的特征空间 P 上, 样本点的协方差矩阵为:

$$C = \frac{1}{N} \sum_{k=1}^N \Phi(x_k) \Phi^T(x_k) \quad (3)$$

这时, 在高维特征空间 P 上的 PCA 就是求解 $\lambda^{\Phi} v = C v$, 即特征值 λ 和特征向量 v 。在高维特征空间 P 中逐个将样本和 $\lambda^{\Phi} v = C v$ 做内积如下:

$$\lambda [\Phi(x_k) v] = \Phi(x_k) C v \quad (k = 1, 2, \dots, N) \quad (4)$$

假设在高维特征空间 P 上, 特征向量 v 可由下式线性表示出, 即:

$$v = \sum_{k=1}^N \alpha_k \Phi(x_k) \quad (5)$$

从上述几个式子合并得出:

$$\lambda^{\Phi} \sum_{j=1}^N \alpha_j \langle \Phi(x_i), \Phi(x_j) \rangle = \frac{1}{N} \sum_{j=1}^N \alpha_j \sum_{k=1}^N \langle \Phi(x_i), \Phi(x_k) \rangle \langle \Phi(x_k), \Phi(x_j) \rangle \quad (i = 1, 2, \dots, N) \quad (6)$$

定义 $N \times N$ 核矩阵 $K = (K_{ij})$,

$$K_{ij} = \langle \Phi(x_i), \Phi(x_j) \rangle \quad (7)$$

上述式子可以简化为

$$N \lambda^{\Phi} K \alpha = K^2 \alpha \quad (8)$$

因为

$$N \lambda^{\Phi} K = K \alpha \quad (9)$$

为了提取主要的特征, 需要对测试样本在高维特征空间 P 上的向量 $v^{(k)}$ 上的投影系数是:

$$\chi_k = \langle \psi^{(k)}, \Phi(x) \rangle = \sum_{j=1}^N \alpha_j^k \langle \Phi(x_j), \Phi(x) \rangle \quad (k=1, 2, \dots, m) \quad (10)$$

上述式子是在理想的环境即假设数据均值是零,但是一般来说,中心化是不成立的. 因此,不能先中心化,需要:

$$\bar{K} = K - I_N K - K I_N + I_N K I_N \quad (11)$$

基于 ReliefF 优化的核主成分分析的二次特征选择算法的输入、输出和执行过程如下.

输入: 过程重复次数 m , 训练样本集 OS , 特征的权重阈值 δ .

输出: 最优特征子集 FS 并且每个特征的权重是大于 δ .

利用可分离性判据:

$$J(j) = \frac{S_b(j)}{S_w(j)} \quad (12)$$

其中 $S_b(j)$ 是代表总的类间离差矩阵; $S_w(j)$ 代表总的类内离差矩阵. $J(j)$ 越大代表类的可分离性就越好,那么就降低了原始数据的分类难度.

调用 Relieff 算法对所有样本进行处理.

输入数据 $S = \{x_1, x_2, \dots, x_l\}$, 投影后数据的维度是 k ; 计算核矩阵:

$$K = (k_{ij})_{l \times l}, k_{ij} = K(x_i, x_j), i, j = 1, 2, \dots, l \quad (13)$$

将核矩阵中心化:

$$K \leftarrow K - \frac{1}{l} \mathbf{j} \mathbf{j}^T K - \frac{1}{l} K \mathbf{j} \mathbf{j}^T + \frac{1}{l^2} (\mathbf{j}^T K \mathbf{j}) \mathbf{j} \mathbf{j}^T$$

向量的元素是 1;

计算上述核矩阵的特征值和特征向量:

$$[\lambda, \Lambda] = \text{eig}(K) \quad (14)$$

其中 eig 为计算矩阵 K 特征值和特征向量的函数.

将特征向量规范化:

$$\alpha_i = \alpha_i / \sqrt{\lambda_i}, i = 1, 2, \dots, k \quad (15)$$

将数据重建:

$$s_i(i) = \sum_{j=1}^l \alpha_j(i) K(x_j, x_i), i, j = 1, 2, \dots, k \quad (16)$$

$$S = \{s_1, s_2, \dots, s_l\} \quad (17)$$

For $A = 0, A < N$

$$\text{merit}[N] = nW[A] / \sqrt{n + n(n-1)r_f N_f} \quad (18)$$

虽然 Relief 算法只能够删除对区分类别能力不强的特征. 但是加入了去除冗余的算法后,当有些特征与分类没有关系时,也就是和类别不相关时,这些特征就会被删除,因为这些特征的权重是不会随着过程重复的次数而变化的. 所以优化之后的特征选择算法既可以去除无相关的特征也可以去除冗余的特征还能够精简数据的维度.

2 实验数据分析

表 1 实验所用数据

数据名字	实例个数	属性 (不包括类别)	类数
glass	214	9	7
balloons	76	4	2
iris	153	4	3
Vote	435	16	2

实验是在 Weka^[5] 上完成,表 1 中的数据是 Weka 平台数据库中自带的数据库. 专门用于算法实验中. Weka 是用于机器学习^[6-17] 算法和数据预处理^[18] 的工具. 数据集^[19] 选自 UCI 数据集,在试验中我们选用的是 k-最近邻^[20] 分类算法(kNN). 首先由训练数据构建分类模型,再将模型用于测试数据并且预测未知类别的对象的所属类别. 再计算出分类的正确率,通过分类的正确率来判断在特征选择上所选用的算法是否最优.

试验采用的四组数据有 glass 数据,含有 214 个实例样本,实例中包含 7 个类别, glass 数据共有 9 个特征. balloons 数据中含有 76 个实例样本,实例中包含 2 个类别, balloons 数据共有 4 个特征. iris 数据中含有 153 个实例样本,实例中包含 3 个类别, iris 数据共有 4 个特征. Vote 数据含有 435 个实例样本,实例中包含 2 个类别, Vote 数据共有 2 个特征.

图 2 和图 3 分别是运用 ReliefF-PCA 得到的 Vote 分类结果散点图和运用基于优化的 ReliefF-KPCA 算法得到的 Vote 分类结果散点图,图 3 比图 2 关于 democrat 类和 republican 的正方形代表的错误分类点明显减少,并且正确分类的点聚集的更紧密,代表分效果更好.

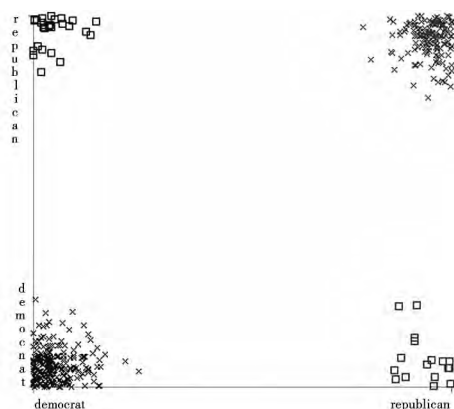


图 2 运用 ReliefF-PCA 得到的 Vote 分类结果散点图

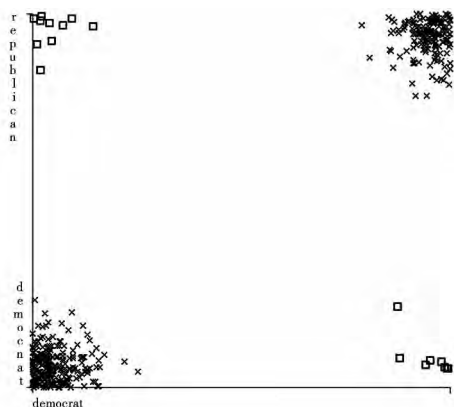


图 3 运用基于优化的 ReliefF-KPCA 算法得到的 Vote 分类结果散点图

同理我们对实验其他 3 组数据进行测试,得到表 2.

表 2 KPCA 特征值的累计贡献率

分类正确率	ReliefF-PCA 算法 /%	基于优化的 ReliefF-KPCA 算法 / %
glass	92.658	94.682
balloons	89.276	92.754
iris	96	98.231
Vote	91.787	95.858

通过以上所有实验结果的对比分析,可以得出 ReliefF-KPCA 的二次特征选择算法的分类正确率是最高的.

3 结 语

本文提出的 ReliefF-KPCA 的二次特征选择算法能够处理高维度、包含无关和冗余特征的数据,提高了算法通用度. 不过对于小样本事件和含有噪音的数据的分类性能不是很高. 需要结合实际情况研究新的特征选择算法.

参 考 文 献:

[1] 刘依恋. 模式分类中特征选择算法研究 [D]. 哈尔滨: 哈尔滨

理工大学, 2014.

- [2] 陆景辉. 基于信息理论的特征选择算法研究 [D]. 北京: 北京交通大学, 2007.
- [3] 李晓岚. 基于 Relief 特征选择算法的研究与应用 [D]. 大连: 大连理工大学, 2013.
- [4] 刘华文. 基于信息熵的特征选择算法研究 [D]. 长春: 吉林大学, 2010.
- [5] 王新莹. 最优特征选择算法在文本分类上的应用研究 [D]. 成都: 电子科技大学, 2013.
- [6] CAO Longbing, ZHANG Huaifeng, ZHAO Yanchang, et al. Combined Mining: Discovering Informative Knowledge in Complex Data [J]. IEEE, 2011, 3(41): 699-712.
- [7] 廉捷. 基于用户特征的社交网络数据挖掘研究 [D]. 北京: 北京交通大学, 2014.
- [8] 王丽文. 基于社交网络的数据挖掘研究 [D]. 西安: 西安电子科技大学, 2014.
- [9] 刘伟, 张凤荔, 程红蓉, 等. 改进 ReliefF 算法在图像型垃圾邮件检测中的应用研究 [J]. 计算机应用研究, 2009(9): 3256-3258, 3262.
- [10] 李玉珍, 王宜怀. 主成分分析及算法 [J]. 苏州大学学报: 自然科学版, 2005(1): 32-36.
- [11] 李靖. 基于改进的主成分分析类算法的人脸识别 [D]. 天津: 天津大学, 2012.
- [12] 朱述龙. 快速近似主成分分析算法 [J]. 遥感学报, 1999(1): 44-48.
- [13] 毛勇, 周晓波, 夏铮, 尹征, 孙优贤. 特征选择算法研究综述 [J]. 模式识别与人工智能, 2007(2): 211-218.
- [14] 姜百宁. 机器学习中的特征选择算法研究 [D]. 北京: 中国海洋大学, 2009.
- [15] 王淑珍. 机器学习算法的 Weka 嵌入 [D]. 广州: 华南理工大学, 2013.
- [16] 李运. 机器学习算法在数据挖掘中的应用 [D]. 北京: 北京邮电大学, 2015.
- [17] 何清, 李宁, 罗文娟, 史忠植. 大数据下的机器学习算法综述 [J]. 模式识别与人工智能, 2014(4): 327-336.
- [18] 王继魁, 吕凯, 李虹. 基于决策树分类的 Weka 平台上数据挖掘应用 [J]. 白城师范学院学报, 2013(5): 37-40.
- [19] 廖强. 基于关联规则挖掘的 weka 数据挖掘应用 [J]. 计算机光盘软件与应用, 2012(19): 52-53.
- [20] 张庆国, 张宏伟, 张君玉. 一种基于 k 最近邻的快速文本分类方法 [J]. 中国科学院研报, 2005, 22(5): 554-559.

(编辑: 温泽宇)