基于信息增益和基尼不纯度的 K 近邻算法

孙 傲,赵礼峰
(南京邮电大学 理学院,江苏 南京 210023)

摘 要:传统 K 近邻算法忽略每个属性对分类的不同重要程度,将每个属性同等看待,在计算样本间距离时赋予每个属性相同的权重,影响样本分类的正确性。利用单一指标来确定属性重要性过于片面,无法全面反应属性对分类的重要程度。针对这一问题,利用信息增益和基尼不纯度的综合指标作为判断属性重要程度的指标,该综合指标越大,属性对分类的重要程度越高。并依据综合指标构造属性权重,计算样本间的加权距离进行分类。为验证该方法的有效性,分别基于 UCI 数据库中 Iris 数据集和 Wine 数据集对基于信息增益和基尼不纯度综合指标的加权 K 近邻算法进行仿真实验,并与传统 K 近邻算法和基于信息增益加权 K 近邻算法进行对比,基于信息增益和基尼不纯度综合指标的加权 K 近邻算法错误率均低于传统 K 近邻算法和基于信息增益加权 K 近邻算法。结果表明该方法比传统 K 近邻法和基于单一指标加权 K 近邻算法能更有效地对样本进行分类。

关键词:数据挖掘;K 近邻;信息增益;基尼不纯度
中图分类号:TP301.6 **文献标识码**:A **文章编号**:1673-629X(2019)09-0051-04
doi:10.3969/j.issn.1673-629X.2019.09.010

K-Nearest Neighbor Algorithm Based on Information Gain and Gini Impurity

SUN Ao,ZHAO Li-feng
(School of Science,Nanjing University of Posts and Telecommunications,Nanjing 210023,China)

Abstract:The traditional K-nearest neighbor algorithm ignores the importance of each attribute to the classification, and treats each attribute equally. When calculating the distance between samples,the same weight is given to each attribute,which affects the correctness of the sample classification. The use of a single indicator to determine the importance of attributes is too one-sided and does not fully reflect the importance of attributes to classification. Aiming at this problem, the comprehensive index of information gain and Gini impurity is used as the index to judge the importance of the attribute. The larger the comprehensive index,the higher the importance of the attribute to the classification. The attribute weights are constructed according to the comprehensive index,and the weighted distance between the samples is calculated for classification. In order to verify the effectiveness of the proposed method,the weighted K-nearest neighbor algorithm based on information gain and Gini impurity comprehensive index is simulated based on Iris dataset and Wine dataset in UCI database,and compared with traditional K-nearest neighbor algorithm and information gain-based weighting. Compared with the K-nearest neighbor algorithm, the error rate of the weighted K-nearest neighbor algorithm based on the information gain and Gini integrity comprehensive index is lower than the traditional K-nearest neighbor algorithm and the information-gain-weighted K-nearest neighbor algorithm. The results show that the proposed method can classify samples more effectively than the traditional K-nearest neighbor method and the single-index weighted K-nearest neighbor algorithm.

Key words:data mining;K-nearest neighbor;information gain;Gini impurity

1 概 述

分类是数据挖掘领域的一个重点和热点方向。分类的方法众多,如决策树、K 近邻算法、遗传算法、支持向量机、神经网络等。其中 K 近邻算法以其理论简单有效,且易于实现等优点在数据挖掘和机器学习领域被广泛应用。

K 近邻算法的直观思想就是给定训练数据集,选择恰当的距离函数,通过已选的距离函数计算待分类

收稿日期:2018-10-16 **修回日期**:2019-02-19 **网络出版时间**:2019-04-24
基金项目:国家自然科学基金青年基金项目(61304169)
作者简介:孙 傲(1993-),男,硕士研究生,研究方向为信息统计与数据挖掘;赵礼峰,教授,硕导,研究方向为图论及其在通信中的应用。
网络出版地址:http://kns.cnki.net/kcms/detail/61.1450.TP.20190424.1047.034.html

样本和训练集中每个样本之间的距离,选取与待分类样本距离最小的 k 个样本作为待分类样本的 k 个最近邻,将待分类样本归为 k 个最近邻所属类别中的多数类。

K 近邻算法虽然算法直观、简单有效,但是也存在着比较明显的缺点。首先 k 值的选择会直接影响分类的效果,k 值较小易导致过拟合,此时待分类样本的类别判断只受与待分类样本较近的实例影响;k 值较大易导致学习的近似误差增大,距离待分类样本较远的实例也会对分类结果产生影响。其次,K 近邻算法在解决大规模数据分类时,分类速度较慢,其需要存储数据集中所有数据样本,计算待分类样本与数据集中所有样本之间的距离。另外,将每个属性等同看待,赋予相同权重,忽略了每个属性对分类的不同重要程度,影响了分类结果的准确性。

针对 K 近邻算法的不足,许多学者对其进行了研究和改进,逐渐完善了 K 近邻算法体系。针对 k 值选择影响分类结果的问题,路敦利等[1]将 BP 神经网络与 kNN 相结合,通过对训练样本使用 k 值不同的 K 近邻算法进行初步分类,同一数据会得到多个不同的初步分类结果集;然后将初步分类结果集作为 BP 神经网络的输入,再对 BP 神经网络进行训练分类。优化了传统 K 近邻算法精度受 k 值影响的问题,提高了分类准确率。孙可等[2]提出 SA-KNN 算法,优化传统 K 近邻 k 值选择问题,引入系数学习理论,通过局部保持投影 LPP 重构测试样本,利用 l2,1 范数去除噪声样本,寻找投影变换矩阵进而确定 k 值。豆增发等[3]提出一种根据最小距离个数来增长式地确定 k 值的方法,将待分类样本与所有训练样本的欧氏距离按大小排序,从最小距离开始,按距离大小等级得到 m 个 k 值选择序列,选择使样本正确分类的最小 k 值,作为 K 近邻的 k 值;同时运用信息增益确定属性重要程度,计算加权距离进行 K 近邻分类。

属性等同看待忽视了属性对分类的重要性,王增民等[4]根据信息熵理论,计算各分类属性与分类的相关性,剔除相关度低的样本,并依据保留特征的相关性作为权重,计算加权欧氏距离,选择 K 近邻,优化了距离度量。陈振洲等[5]提出 FWKNN 算法,将 SVM 算法应用到 K 近邻特征权重的度量中,以 SVM 算法中参数 w 作为特征的权重,并以此计算加权欧氏距离,选取 k 个近邻。

合适的距离度量函数影响分类效果,周靖等[6]将样本特征参数的熵值与样本分布概率的乘积作为特征对分类的相关度,并根据相关度值衡量特征对分类影响程度的强弱以定义样本间的距离函数,优化了在进行近邻选择时多数类别和高密度样本占优的情况。杨立等[7]提出 SDKNN 算法,分析属性内取值的语义差异,定义语义距离构成距离函数,优化了同一属性取值的语义差异所带来的影响。

分类速度随数据维度和样本量的增加急剧下降,余小鹏等[8]针对 K 近邻算法搜索慢的缺陷,提出自适应 K 最近邻算法,建立半径生长的 BP 神经网络模型,在以测试点为中心的超球内搜索,缩小了搜索范围,提高了搜索效率。江昆等[9]提出运用随机森林算法对变量进行排序,剔除不重要的变量,降低数据维数,提高 K 近邻算法在处理高维数据集的性能。Chen Yewang 等[10]针对 K 近邻分类速度慢的问题,提出一种改进基于缓冲区的 K 最近邻查询算法,有效减少搜索待分类样本的时间复杂度。

不同的决策方式使 k 个近邻对待分类样本有不同的分类,杨金福等[11]在模板约简 K 近邻算法的基础上提出 TWKNN 算法,利用模板约简技术,将训练集中远离分类边界的样本去掉,同时按照各个近邻与待测样本的距离为 k 个近邻赋予不同的权值,增强了算法的鲁棒性。肖辉辉等[12]提出一种 FCD-KNN 算法,首先通过距离度量函数计算待分类样本与训练样本的距离值,通过距离值选择 k 个样本,定义类可信度函数,根据各类近邻样本点的平均距离及个数判断待测试样本的类别,优化了 K 近邻决策的方式。

但是,目前对于 K 近邻算法的改进和应用[13-17],大多使用不加权的欧氏距离或单一指标的加权欧氏距离,无法全面准确地度量分类特征对分类的重要程度。针对这一问题,利用信息增益和基尼不纯度的综合指标确定分类属性的重要程度,只有信息增益足够大且基尼不纯度足够小的特征才是真正重要的特征,即用信息增益和基尼不纯度的差值作为衡量特征重要程度的依据,并根据此综合指标构造权重,计算加权欧氏距离,进行 K 近邻分类。

2 相关知识

2.1 K 近邻算法

K 近邻算法是一种基于实例的懒散算法,其思想就是选择距离度量函数,根据距离度量,找出与待分类样本距离最小的 k 个最近邻,通过多数表决等方式进行类别预测。

算法流程如下:

(1)选择距离度量函数,计算待分类样本与数据集中所有实例的距离值;

(2)在数据集实例中,找出与待分类样本距离度量值最小的 k 个实例点;

(3)根据与待分类样本距离度量值最小的 k 个实例点所属的类别,通过分类决策规则(如多数表决),

决定待分类样本的类别。

不同的距离度量方式,所选取的 k 个最近邻不同,从而最终待分类样本的类别判断也会不同。因此选取合适的距离度量方式,对提高 K 近邻分类的性能尤为重要。

2.2　信息增益及算法

信息增益表示在已知某特征的信息而使得类别的信息的不确定性减小的程度。若特征为 A,训练数据集为 D,则特征 A 对训练数据集 D 的信息增益为 $g(D,A)$。

$$g(D,A) = H(D) - H(D \mid A) \tag{1}$$

其中,$H(D)$ 为集合 D 的经验熵,表示对数据集 D 进行分类的不确定性;$H(D \mid A)$ 为在特征 A 给定条件下数据集 D 的经验条件熵,表示在给定特征 A 给定条件下对数据集 D 进行分类的不确定性。

因此信息增益就表示在给定特征 A 而使得对数据集 D 进行分类的不确定性减少的程度,所以信息增益大的特征具有更强的分类能力,该特征对分类的重要程度更大。

信息增益的算法流程如下:

(1)计算数据集 D 的经验熵 $H(D)$。

$$H(D) = -\sum_{k=1}^{k} \frac{|C_k|}{|D|} \log_2 \frac{|C_k|}{|D|} \tag{2}$$

其中,$|C_k|$ 为属于类 C_k 的样本个数;$|D|$ 为样本容量。

(2)计算特征 A 对数据集 D 的经验条件熵 $H(D \mid A)$。

$$H(D \mid A) = -\sum_{i=1}^{n} \frac{|D_i|}{|D|} \sum_{k=1}^{k} \frac{|D_{ik}|}{|D_i|} \log_2 \frac{|D_{ik}|}{|D_i|} \tag{3}$$

其中,$|D_i|$ 为根据特征 A 的取值将 D 划分的第 i 个子集的样本个数;$|D_{ik}|$ 为子集 D_i 中属于类 C_k 的样本个数。

(3)计算信息增益。

2.3　基尼不纯度

基尼不纯度表示集合的不确定性,基尼不纯度越大,集合的不确定性也就越大。若经过某特征对集合进行划分,使得划分后集合的不确定性减小,则该特征较重要。

在分类问题中,假设集合 D 有 k 个类,则集合 D 的基尼不纯度为:

$$\mathrm{Gini}(D) = 1 - \sum_{k=1}^{k} \left(\frac{|C_k|}{|D|}\right)^2 \tag{4}$$

其中,C_k 是 D 中属于第 k 个类的样本子集;k 是类的个数。

经过特征 A 划分后,集合 D 的基尼不纯度为:

$$\mathrm{Gini}(D,A) = \sum_{i=1}^{n} \frac{|D_i|}{|D|} \mathrm{Gini}(D_i) \tag{5}$$

其中,D_i 为集合 D 中属性 A 取值为 i 的样本子集。

3　基于信息增益和基尼不纯度综合指标的加权 K 近邻分类器构造方法

基于信息增益和基尼不纯度综合指标的加权 K 近邻分类器将信息增益和基尼不纯度两个指标结合起来评价属性的重要性,综合利用了信息增益和基尼指数的两个评价指标的优点,对属性重要性的评价更加全面。信息增益越大的特征重要程度越高,基尼不纯度越小的特征重要程度越高。由此出发,若一个特征基尼不纯度足够小且信息增益足够大即信息增益和基尼不纯度的差值越大,则表示该特征对分类越重要。

构造流程如下:

(1)将信息增益和基尼不纯度综合指标定义为 GaGi,属性 A 的 GaGi 指标的计算公式为:

$$\mathrm{GaGi}(A) = g(D,A) - \mathrm{Gini}(D,A) \tag{6}$$

计算数据集每个属性的 GaGi 指标,信息增益越大,属性越重要;基尼不纯度越小,属性越重要。所以属性的综合指标 GaGi 的数值越大,该属性的重要程度越大。

(2)根据属性的 GaGi 指标值构造属性权重,训练数据集 D 共有 n 个属性,其中属性 A 的权重构造公式为:

$$\mathrm{weight}(A) = \frac{\mathrm{GaGi}(A)}{\sum_{i=1}^{n} \mathrm{GaGi}(i)} \tag{7}$$

(3)根据每个属性的权重,建立加权欧氏距离 d,样本点 x_1, x_2 之间的加权欧氏距离为:

$$d(x_1,x_2) = \sqrt{\sum_{i=1}^{n} \mathrm{weight}(i)(x_1^i - x_2^i)^2} \tag{8}$$

(4)计算待分类样本与所有训练样本的加权欧氏距离,对距离值进行排序。

(5)指定 k 值,选取与待分类样本最近邻的 k 个训练样本。

(6)依据多数表决决策规则,将 k 个最近邻样本所属类别中的多数类别作为待分类样本的类别。

4　仿真实验

4.1　实值仿真

利用 Python3 编程环境构造分类器,分别采用 UCI 数据库中的 Iris 和 Wine 两个数据集验证该算法的有效性。为保证有充足的数据集用于训练,采用留存交叉验证的方法对每个数据集随机选择 20% 的数据作为测试集,剩下的作为训练集。为更加准确地估计分

类器的错误率,进行多次迭代求出平均错误率作为分类器的最终错误率。

数据集信息如表 1 所示。

表 1 数据集信息

数据集	特征数	类别数	实例数
Iris	4	3	150
Wine	13	3	178

4.1.1 Iris 数据集仿真效果

该数据集包含 150 个实例和 4 个特征变量,特征变量分别为:花萼长度 sepal length、花萼宽度 sepal width、花瓣长度 petal length、花瓣宽度 petal width。

选取 30 个数据实例作为测试集,剩下 120 个数据实例作为训练集,用于构造加权分类器,同时为避免特征的数量值大小对样本间距离的影响,消除特征间数量值等级的差别,对数据集进行归一化处理。最后采用 10 次留存交叉验证法得到最终分类器错误率。

在不同 k 值下算法实现的平均错误率如表 2 所示。

表 2 不同 k 值下的三种算法错误率对比(Iris)

k	传统 K 近邻算法	基于信息增益的加权 K 近邻算法	基于信息增益和基尼不纯度综合指标的加权 K 近邻法
1	0.047	0.033	0.040
2	0.057	0.043	0.030
3	0.047	0.033	0.030
4	0.040	0.043	0.027
5	0.037	0.043	0.023

通过对比可以发现,加权 K 近邻算法总体优于传统 K 近邻算法,属性权重的度量方式和 k 值的选择不同也会影响分类精度,在此数据下,总的来说基于信息增益和基尼不纯度综合指标的加权 K 近邻法更优。

4.1.2 Wine 数据集仿真效果

UCI 数据库中 Wine 数据集包含 178 个数据实例,有 13 个红酒理化元素特征变量,对数据集进行归一化,随机选择 35 个数据实例作为测试集,剩下 143 个数据作为训练集,最后采用 10 次留存交叉验证法得到最终分类器错误率。

k 值取 1 到 5 时,三种 K 近邻算法的平均错误率如表 3 所示。

通过对 Wine 数据集的算法仿真可以看出,基于信息增益和基尼不纯度综合指标的加权 K 近邻法最优,加权 K 近邻算法始终优于传统 K 近邻算法,在 k 等于 1 和 5 时分类器的准确度最高,分类效果最好。

表 3 不同 k 值下的三种算法错误率对比(Wine)

k	传统 K 近邻算法	基于信息增益的加权 K 近邻算法	基于信息增益和基尼不纯度综合指标的加权 K 近邻法
1	0.043	0.043	0.014
2	0.049	0.049	0.017
3	0.040	0.034	0.017
4	0.029	0.026	0.017
5	0.043	0.034	0.014

4.2 方法评价与不足

通过对两个真实数据集的仿真实验,可以得出区别对待属性特征,进行加权 K 近邻算法明显优于传统的 K 近邻方法;同时权重的构造指标也会对分类结果产生影响。实验证明,采用信息增益与基尼不纯度的综合指标比采用单一的信息增益指标划分属性重要程度更加合理,分类精度略优于采用单一的信息增益指标的 K 近邻法。

该算法也存在一定的不足,信息增益和基尼不纯度偏向于取值较多的特征。在特征取值较均衡时,算法具有很好的分类性能,在特征取值严重不均衡时,算法分类性能有所下降。

5 结束语

合理划分分类属性的重要程度,是数据挖掘分类问题和特征选取的重点方向。面对数据集的众多分类特征,如何选择合适指标评价特征变量的重要程度至关重要。针对传统 K 近邻算法将特征变量等同权重影响分类准确性和单一指标度量特征变量重要程度的不全面性的问题,提出一种基于信息增益和基尼不纯度的加权 K 近邻算法。该算法通过定义的 GaGi 指标对分类特征的重要程度进行度量,并以此为权重,计算加权欧氏距离,进行 K 近邻分类。对 UCI 数据库中 Wine 和 Iris 两个数据集的实验结果表明,该方法分类错误率比传统 K 近邻算法和基于信息增益的加权 K 近邻算法的错误率更低,分类性能更好。

参考文献:

[1] 路敦利,宁 芊,臧 军. 基于 BP 神经网络决策的 KNN 改进算法[J]. 计算机应用,2017,37(S2):65-67.

[2] 孙 可,龚永红,邓振云. 一种高效的 K 值自适应的 SA-KNN 算法[J]. 计算机工程与科学,2015,37(10):1965-1970.

[3] 豆增发,王英强,王保保. 一种基于信息增益的 K-NN 改进算法[J]. 电子科技,2006(12):52-56.

[4] 王增民,王开珏. 基于熵权的 K 最临近算法改进[J]. 计算机工程与应用,2009,45(30):129-131.

(下转第 116 页)