基于引力模型的朴素贝叶斯分类算法*

王 威†, 赵思逸, 王 新
(长沙理工大学 计算机与通信工程学院, 长沙 410114)

摘 要: 针对朴素贝叶斯分类器在分类过程中不同类别的同一特征量之间由于存在相似性,易导致误分类的现象,提出基于引力模型的朴素贝叶斯分类算法。提出以引力公式中距离变量的平方作为相似距离,应用引力模型来刻画特征与其所属类别之间的相似度,从而克服朴素贝叶斯分类算法容易受到条件独立假设的影响而将所有特征同质化的缺点,并能有效地避免噪声干扰,达到修正先验概率、提高分类精度的目的。对遥感图像的分类实验表明,基于引力模型的朴素贝叶斯分类算法易于实现、可操作性强,且具有更高的平均分类准确率。

关键词: 分类算法; 朴素贝叶斯; 引力模型; 遥感图像
中图分类号: TP391; TP301.6 **文献标志码**: A **文章编号**: 1001-3695(2018)09-2602-03
doi:10.3969/j.issn.1001-3695.2018.09.009

Naive Bayesian classification algorithm based on gravity model

Wang Wei†, Zhao Siyi, Wang Xin
(*School of Computer & Communication Engineering, Changsha University of Science & Technology, Changsha* 410114, *China*)

Abstract: In order to solve the problem of misclassified in the process of naive Bayesian classifier which caused by the similarity between the same feature quantities of different categories, this paper presented a simple Bayesian classification algorithm based on gravitational model. This algorithm could overcome the influence of the naive Bayesian classification algorithm, which easy to be influenced by effectively avoid noise interference, correct the prior probabilities, and could improved the accuracy of classification purposes. This paper proposed a gravitational model to describe the similarity between the feature and its category by using the square of the distance variable in the gravitational formula as the similar distance. The classification experiments of remote sensing images show that the naive Bayesian classification algorithm based on gravitational model is easy to implement, has high operability and has higher average classification accuracy.

Key words: classification algorithm; naive Bayesian; gravitational model; remote sensing image

0 引言

随着互联网的迅速发展,人们接收信息的途径也越来越多,每天能接收到的信息呈爆炸式增长。如何快速准确地对这些信息进行分类,方便人们阅览、查找,是一个重要课题。

贝叶斯(Bayesian)算法[1]发源于古典数学理论,以贝叶斯定理为基础,假设待分类样本遵循某种概率分布,根据已观测样本数据对测试样本的类别进行概率计算,从而得出的优分类决策。贝叶斯分类算法逻辑清晰、算法简单,能够处理同类型的数据,所以应用十分广泛。但是贝叶斯算法在处理多层逻辑结构数据时,算法的复杂度过高,影响分类效率。朴素贝叶斯(navie Bayesian,NB)算法[2]是从贝叶斯分类算法中发展而来的,是贝叶斯分类算法之一。朴素贝叶斯算法结构较简单,模型对缺失数据不太敏感,计算过程简单高效易于实现,在大量文本集上分类效果甚至超过支持向量机(support vector machine, SVM)[3]、人工神经网络(artificial neural network, ANN)[4]等分类算法[5]。但是朴素贝叶斯算法假设属性之间相互独立[6],这在实际应用中通常都无法满足,导致其在某些方面的分类效率不高。虽然如此,朴素贝叶斯分类算法还是以高精度和高效率以及最小的误分类率等优点[1]取得了较大的成功,是目前较常使用的分类算法之一。

为了克服条件独立假设带来的缺陷,提高贝叶斯分类算法的效率,学者们提出了基于属性加权的朴素贝叶斯[7,8]。属性加权是指根据属性对分类的贡献程度给不同的属性分配不同的权值,另一方面降低了属性之间的依赖关系,即放松了属性之间的条件独立假设,同时保证了朴素贝叶斯的分类精确度和运行时间。为了克服条件独立假设对朴素贝叶斯的限制,同时为了避免冗余数据、噪声对算法分类性能的影响,周喜[9]结合粗糙集的属性约简算法与朴素贝叶斯分类算法,提出了基于核属性的加权朴素贝叶斯的分类模型,一方面使给定数据得到了一定程度的简化,一方面可以通过改变权值的方法放松条件独立假设,减小了冗余属性对分类的影响,使分类精度得到提高。基于核属性的加权朴素贝叶斯分类模型的基本思想是先对数据集合进行离散化和缺失数据的预处理操作,然后通过粗糙集对属性进行约简操作,最后利用加权的朴素贝叶斯算法进行分类。其中数据的预处理是提高算法效率的重要步骤,但是数据预处理的结果可能会导致一些数据的偏差并影响分类结果,故本算法虽然在分类效率方面表现良好,但扩展性不强。王行甫等人[10]提出了基于余弦相似度和实例加权改进的贝叶斯算法(IWIMNB),使用余弦相似度度量样本的相似性,选出最优训练样本子集,用相似度作为训练样本的权值来训练修正后的贝叶斯模型进行分类,达到了提高分类精度的目的,实验结果证明该算法在小样本文本分类中的分类效率不如 SVM,但是随着样本数量增大,该算法的分类效率得到提升。Shahnaj 等人[11]提出了基于权重矩阵的朴素贝叶斯分类算法,该算法用于文本分类,根据文本的词频为贝叶斯网络分配一个权重矩阵,为训练单词进行评分,以提高分类器的分类效率。

收稿日期: 2017-04-25; **修回日期**: 2017-06-02 **基金项目**: 国家重大基础研究项目(613XXX0301)

作者简介: 王威(1974-),男(通信作者),山东青岛人,教授,博士(后),主要研究方向为智能信息处理(wangwei@csust.edu.cn);赵思逸(1989-),女,湖南永州人,硕士研究生,主要研究方向为遥感图像处理;王新(1976-),女,湖北武汉人,讲师,硕士,主要研究方向为智能信息处理.

本文提出基于引力模型[12]的朴素贝叶斯分类算法(naive Bayesian based on gravity model,G-NB),利用经典物理学中的万有引力模拟朴素贝叶斯分类概率。

1　基于引力模型建模

1.1　朴素贝叶斯

贝叶斯网络(Bayesian network,BN)又称为信度网络(belief network)[13],是用来描述数据变量之间因果关系的图模型。贝叶斯网络综合考虑先验信息和样本数据进行分析,将变量间的潜在关联用图解模型表达出来,更易于理解。综合来说,贝叶斯网络有强大的知识表达和推理能力,用于因果推理和不确定性知识表达,能够处理不完全的数据[14]。

朴素贝叶斯在贝叶斯网络的基础上提出了一个条件独立假设,大大简化了条件概率的求解难度,因此得到快速推广。条件独立假设认为朴素贝叶斯网络中任意一个节点的概率,在给定父节点的情况下,其概率与非 a_i 的子节点集无关,即

$$p(a_i)\mid A(a_i),P(a_i)=p(a_i)\mid P(a_i) \tag{1}$$

其中:$A(a_i)$ 为任一 a_i 非子孙节点集合概率;$P(a_i)$ 为 a_i 的父节点集合概率;$p(a_i)$ 为节点概率。

假设训练数据集中有 m 个分类标志 $\Omega=\{C_1,C_2,\cdots,C_m\}$,每一条待分类数据共有 n 个属性 $A_1,A_2,\cdots,A_n$,现假定 X 是一个分类标志未知的样本,$X=(x_1,x_2,\cdots,x_n)$,其中第 i 项 x_i 表示的是样本 X 的第 i 个属性 A_i 的取值,于是样本 X 的分类属性取值为 C_i 的概率可由贝叶斯公式来计算:

$$P(X\mid C_i)=\prod_{k=1}^{m}P(x_k\mid C_i) \tag{2}$$

其中:$P(x_k\mid C_i)$ 为样本 X 的第 i 个属性 A_i 的取值 x_i 属于 C_i 的概率。则朴素贝叶斯分类器的计算模型为

$$C(X)=\arg\max_{C_i\in\Omega}P(C_i)\prod_{k=1}^{m}P(x_k\mid C_i) \tag{3}$$

在实际应用中,样本 A_k 的取值类型有连续和离散两种情况。

若样本取值类型为离散的,则属性 x_i 属于 C_i 的概率为

$$P(x_k\mid C_i)=S_{ik}/S_i \tag{4}$$

其中:S_i 为训练数据样本中分类为 C_i 的样本实例数量;S_{ik} 表示在训练数据样本集中分类为 C_i 且属性 A_k 的取值是 x_k 的样本数目。

若样本取值类型为连续的,可以把每一个连续的属性离散化,然后用相应的离散区间替换连续属性值。通过把连续属性转换成序数属性,计算类 C_i 的训练记录落入 X_k 对应区间的比例来估计条件概率 $P(x_k\mid C_i)$。朴素贝叶斯分类算法的提出是为了降低贝叶斯分类器条件概率计算的复杂度,提出了条件属性独立性假设,然而也正是因为这个假设在实际情况中很难被满足,所以其分类性能也被限制。但是大量的国内外研究表明可以通过一些策略对其进行改进,使得在属性之间具有相互依赖关系时朴素贝叶斯模型仍然具有可接受的分类效果。

1.2　基于引力模型的朴素贝叶斯分类算法

在经典物理学中,质点 b 对 a 的引力可表示为

$$F(a,b)=\frac{Gm_am_b}{r_{ab}^2} \tag{5}$$

其中:G 为空间引力常数;r_{ab} 为 a 与 b 的距离。

在朴素贝叶斯分类算法中,如果分类标志和样本看做有质量的物体,则可以通过与经典物理学中的引力公式进行类比来定义分类与属性间的引力。两点间概率越大,则点间引力越大,从而可以将引力强度作为分类概率测度。

如图 1 所示,将引力模型看做一条直线,分类类别均匀分布在直线上,待分类数据视为一个密度不均匀的直杆,与分类类别在同一条直线上。直杆由于密度不均匀,可将其看做多个质量不同的小块组成。不同小块的重心与分类类别间距离不同。

图1　引力模型

样本质量等于各属性质量相加,样本 $A(A_1,A_2,\cdots,A_n)$ 有 n 个属性,故认为样本 $A(A_1,A_2,\cdots,A_n)$ 的质量 $m_A=\sum_{i=1}^{n}m_i$。各属性与同一分类间的引力方向相同,根据力的合成原理,有

$$F_{AC_i}=\sum_{j=1}^{n}A_jC_i=\sum_{j=1}^{n}\frac{G}{r_{ij}^2}M_im_{ij} \tag{6}$$

其中:m_{ij} 为相对分类标志 C_i,样本 $A(A_1,A_2,\cdots,A_n)$ 中属性 A_j 的质量,其值为属性 A_j 对分类标志 C_i 的先验概率。在没有先验知识的情况下,可认为引力常数 G 为 1。由于概率分类模型中不同属性的概率分布不同,所以本模型中将 r_{ij}^2 设置为属性 A_j 与分类标志 C_i 的相似距离。式(6)可写为

$$F_{AC_i}=\sum_{j=1}^{n}A_jC_i=\sum_{j=1}^{n}\frac{1}{r_{ij}^2}M_im_{ij} \tag{7}$$

其中:r_{ij}^2 为属性 A_j 与分类标志 C_i 的相似距离。

分类的基本任务就是要将多种类别的图像区分开来,其主要技术路线是提取待分类图像特征,所提取的特征设为 N 类,则每一类特征存在 M 种可能性取值,M 由待分类图像的特点决定。由于待分类图像之间存在相似性,所以为了能够显著地将图像区分开,需要采用较好的分类特征。对于一个特征要能有效地刻画其所属的类别,该特征应满足两个方面的要求:a)同一个特征取值的类间方差不应太大,这样方便判断测试数据的值是否属于该特征点的取值范围,同一特征的类内方差越小,则分类越可信;b)两个类间取值差距应较显著,这样更容易分辨测试数据所属类别,若多个分类的取值区间重合程度越高,则其分类越模糊,越不可信。

多个特征联合起来判定分类结果,会使得分类情况更加复杂,因此必须采用有效的度量手段或者方法,将所有的分类结果量化,从而避免误分类。因此提出相似距离来辅助分类。

为了更有效地对数据进行分类,本文采用相似距离 r^2 来度量数据的分类能力。

$$r^2=\sigma^2/d \tag{8}$$

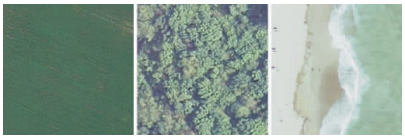
其中:σ^2 为该类特征的测试数据的类内方差;d 是最小类间差。最小类间差即对每一类特征值求平均值,将平均值降序排列后,特征与相邻值差值中较小者。

朴素贝叶斯分类算法中属性的先验概率为 $P(x_k\mid C_i)=S_{ik}/S_i$。其中:$S_i$ 表示训练数据样本中分类表示为 C_i 的样本数量;S_{ik} 表示在训练数据样本集中分类为 C_i 且属性 A_k 的取值是 x_k 的样本实例的数目。这种取值方法仅考虑了样本所占比重,忽略了样本取值变化的梯度。当某一属性中两类分类的取值相近时,这一属性在这两个分类上的分类效率明显有所降低。而相似距离根据测试数据的分布特点,凸出了分类效率高的属性分类强度,削弱了模糊数据对分类的影响,提高了分类效率。

2　结果分析

为了验证本文算法的有效性,从 http://vision.ucmerced.edu/datasets/landuse.html 获取遥感图像作为实验数据,进行验证实验并与传统贝叶斯分类算法(NB)、SVM 算法基于余弦相似度的朴素贝叶斯分类算法(R-NB)进行对比。本文选取了三类

遥感图像,分别是农田、森林、海滩各 100 幅。图像示例见图 2。



(a)农田 (b)森林 (c)海滩

图 2 实验图像示例

从每类图像中选取 30 幅图像用于组成冗余字典,30 幅用于训练,剩下的 40 幅图像中随机选取 20 幅和 40 幅用来作测试。将训练图像和测试图像分别在冗余字典上进行稀疏表示,得到图像的稀疏表示特征向量。将特征向量输入分类器中进行分类。分类特征从特征向量中提取,提取向量的质心位置和最大值位置。为了防止噪声带来的干扰,采用降序法排序后,取前三个值的位置,分别记为第一特征点、第二特征点和第三特征点,代替最大值位置,从而降低噪声的干扰,提高置信度。

使用朴素贝叶斯分类算法、SVM 算法、基于余弦相似度和实例加权改进的贝叶斯算法(IWIMNB)与本文提出的基于引力模型的朴素贝叶斯分类算法(G-NB)分别对数据进行分类。分类正确率对比如表 1 所示。

表 1 不同算法在不同样本数量下的识别率

测试样本数	类别数	NB/%	SVM/%	IWIMNB/%	G-NB/%
60	3	78.33	85.00	80.00	83.33
120	3	80.83	86.67	84.16	87.50
均值	3	79.58	85.84	82.08	85.42

从表 1 可以看出,在不同测试样本数的情况下,G-NB 有着明显高于 NB 和 R-NB 的分类效率,且测试样本数越多,其效率越高。在样本数量较少的情况下,G-NB 的分类效率与 SVM 差不多,但是由于 SVM 算法对大样本数据处理能力不强,从实验数据可看出,随着样本数量增大,G-NB 的分类效率提高度明显高于 SVM,且 G-NB 还可以结合其他属性加权方法,继续提高效率,故本文提出的基于引力模型的朴素贝叶斯分类算法可操作性强,且能显著提高分类效率。

3 结束语

本文提出的基于引力模型的朴素贝叶斯算法,主要是根据训练数据本身的特点,考虑了不同特征对不同分类的辨识度不同,提出了相似距离来减少辨识度低的特征对分类的影响,降低了相似距离较大的分类概率对分类的影响,提高有效分类概率的分类影响,从而提高朴素贝叶斯分类算法的分类效率。对部分图像的稀疏表示系数进行分类的实验结果表明,本文提出的改进算法可操作性强,且能有效提高分类准确率。

参考文献:

[1] Domingos P, Pazzani M. On the optimality of the simple Bayesian classifier under zero-one loss[J]. Machine Learning, 1997, 29(2): 103-130.

[2] McCallum A, Nigam K. A comparison of event models for naive Bayes text classification[C]//Proc of AAAI Workshop on Learning for Text Categorization. 1998: 41-48.

[3] Tarabalka Y, Fauvel M, Chanussot J, *et al*. SVM-and MRF-based method for accurate classification of hyperspectral images[J]. IEEE Geoscience & Remote Sensing Letters, 2010, 7(4): 736-740.

[4] LeDoux J E. Emotion circuits in the brain[J]. Annual Review of Neuroscience, 2000, 23(2): 155-184.

[5] Kulkarni A R, Tokekar V, Kulkarni P. Identifying context of text documents using naïve Bayes classification and Apriori association rule mining[C]//Proc of the 6th International Conference on Software Engineering. Piscataway, NJ: IEEE Press, 2012: 1-4.

[6] 程环环. 基于贝叶斯网络的图像内容表述与分类[D]. 长沙: 国防科学技术大学, 2011.

[7] 李静梅, 孙丽华, 张巧荣, 等. 一种文本处理中的朴素贝叶斯分类器[J]. 哈尔滨工程大学学报, 2003, 24(1): 71-74.

[8] 李方, 刘琼荪. 基于改进属性加权的朴素贝叶斯分类模型[J]. 计算机工程与应用, 2010, 46(4): 132-133.

[9] 周喜. 基于粗糙集的加权朴素贝叶斯分类算法研究[D]. 长沙: 长沙理工大学, 2013.

[10] 王行甫, 付欢欢, 王琳. 基于余弦相似度和实例加权改进的贝叶斯算法[J]. 计算机系统应用, 2016, 25(8): 166-170.

[11] Shati S P, Hossain M D, Nadim M, *et al*. Enhancing performance of naïve Bayes in text classification by introducing an extra weight using less number of training examples[C]//Proc of International Workshop on Computational Intelligence. Piscataway, NJ: IEEE Press, 2016: 142-147.

[12] 赵柳. 相对论与引力理论导论[M]. 北京: 科学出版社, 2016.

[13] 李硕豪, 张军. 贝叶斯网络结构学习综述[J]. 计算机应用研究, 2015, 32(3): 641-646.

[14] 慕春棣, 戴剑彬, 叶俊. 用于数据挖掘的贝叶斯网络[J]. 软件学报, 2000, 11(5): 660-666.

(上接第 2591 页)

[6] Tarko A, Azam S, Inerowicz M. Operational performance of alternative types of intersections: a systematic comparison for indiana conditions [J]. Congress Proceedings, 2010, 32(31): 386-391.

[7] Esawey M E, Sayed T. Comparison of two unconventional intersection schemes: crossover displaced left-turn and upstream signalized crossover intersections[J]. Transportation Research Record Journal of the Transportation Research Board, 2007, 2023: 10-19.

[8] Zhao Jing, Ma Wanjing, Head K L, *et al*. Optimal operation of displaced left-turn intersections: a lane-based approach[J]. Transportation Research Part C: Emerging Technologies, 2015, 61(12): 29-48.

[9] 刘秋晨, 张轮, 杨文臣, 等. 城市道路新型连续流交叉口的设计及仿真[J]. 交通信息与安全, 2013, 31(2): 122-127.

[10] Michael G, Bruce P E, Paul W, *et al*. Continuous flow intersection-gaining speed in the United States[N]. CE News, 2006-01-26.

[11] Jagannathan R, Bared J. Design and performance analysis of pedestrian crossing facilities for continuous flow intersections[J]. Transportation Research Record: Journal of the Transportation Research Board, 2005, 1939(1): 133-144.

[12] Coates A, Yi Ping, Liu Peng, *et al*. Geometric and operational improvements at continuous flow intersections to enhance pedestrian safety [J]. Transportation Research Record: Journal of the Transportation Research Board, 2014, 2436: 60-69.

[13] Washington D. Highway capacity manual[J]. Special Report, 2000, 1(1-2): 5-7.

[14] Japan Society of Traffic Engineers. Manual of traffic signal control [M]. Tokyo: Japan Society of Traffic Engineers, 2006.

[15] 冯树民, 裴玉龙. 行人过街延误研究[J]. 哈尔滨工业大学学报, 2007, 39(4): 613-616.

[16] 冯树民, 裴玉龙. 考虑行人过街的两相位交叉口配时优化[J]. 交通运输系统工程与信息, 2009, 9(3): 146-151.

[17] 高利平. 城市道路环境下人行横道处行人与机动车冲突分析与延误建模[D]. 北京: 北京交通大学, 2010.

[18] 钱大琳, 陈小红. 基于行人专用相位的交叉口信号控制优化模型[J]. 中国公路学报, 2013, 26(5): 140-147.

[19] 马万经, 林瑜, 杨晓光. 多相位信号控制交叉口行人相位设置方法[J]. 交通运输工程学报, 2004, 4(2): 103-106.

[20] 孙迪. 行人过街交通行为分析建模[D]. 长春: 吉林大学, 2012.

[21] 卢凯, 胡建伟, 李福樑, 等. 行人斜穿信号交叉口绿波设计及延误模型[J]. 吉林大学学报: 工学版, 2016, 46(6): 1818-1826.

[22] Wikipedia. Continuous-flow intersection [EB/OL]. (2005-09-22) [2017-05-01]. http://en.wikipedia.org/wiki/Continuous-flow-intersection.

[23] Chen Kuanmin, Luo Xiaoqiang, Ji Hai, *et al*. Towards the pedestrian delay estimation at intersections under vehicular platoon caused conflicts[J]. Scientific Research & Essays, 2010, 5(9): 941-947.