

增强型朴素贝叶斯学习

Boosting Naive Bayesian Learning

王实高文 TP18

(中国科学院计算技术研究所 北京 100080)

Abstract In this paper we introduce the Boosting Naive Bayesian Learning. The Naive Bayesian classifier provides a simple approach to discriminate classification learning that has demonstrated competitive predictive accuracy on many learning tasks and it was broadly used in Data Mining field. Boosting combines the weak hypotheses such as Naive Bayesian by summing their probabilistic predictions. On the real world datasets, Boosting Naive Bayesian method has generalized perfect performance than the best published result using any other learning method.

Keywords Data mining, Naive Bayesian, Boosting

朴素贝叶斯是一种分类监督学习方法^[1]。在理论上,应用其前提为例子的属性值独立于例子的分类属性。这个前提在实际应用中过于严格,常常得不到满足,即使是这样,在违反该前提的情况下,朴素贝叶斯学习方法仍然取得了很大的成功。

近来,一种改进的朴素贝叶斯方法,增强(Boosting),受到广泛的关注,AdaBoost方法^[2]是其主要方法。当AdaBoost方法被用于联合几个朴素贝叶斯分类器时,其在数学上等价于一个具有稀疏编码输入,单隐层节点,sigmoid激活函数的反馈型神经网络^[3]。

增强型朴素贝叶斯学习具有很重要的优点:首先,改进后的朴素贝叶斯分类器具有很好的概括能力。在针对真实世界的数据集情况下,该分类器比任何已知的方法包括反向传播学习算法、C4.5决策树等,给出了更好的测试精度。其次,这些分类器的学习效率非常高,给定具有 f 个属性值的 n 个训练例,其学习复杂度是线性的, $O(nf)$,这在所有的学习算法中是最快的,而且使用的常系数也很少。

1 朴素贝叶斯学习

给定一个具有 k 个属性的数据集,这 k 个属性值均为离散值,分类任务是预测测试集中每一个例子的类别。给定一个具体的例子,其属性值从 a_1 到 a_k ,该例子属于某一个类 c_i 的概率是 $\Pr(C=c_i | A_1=a_1 \wedge \dots \wedge A_k=a_k)$ 。显然如果该例子属于的某一个类的该概率值具有最大值,那么该例子就属于这个类。

根据贝叶斯定理:

$$\Pr(C=c_i | A_1=a_1 \wedge \dots \wedge A_k=a_k) =$$

• 46 •

$$\frac{\Pr(A_1=a_1 \wedge \dots \wedge A_k=a_k | C=c_i) \Pr(C=c_i)}{\Pr(A_1=a_1 \wedge \dots \wedge A_k=a_k)} \Pr(C=c_i) \quad (1)$$

其中, $\Pr(C=c_i)$ 被称为背景概率,可以很容易从训练数据集中得到。 $\Pr(A_1=a_1 \wedge \dots \wedge A_k=a_k)$ 与任何一个 c_i 都无关。这两个参数可以被认为是常量。

如果属性值是独立的话,那么公式(1)可被写作:

$$\Pr(C=c_i | A_1=a_1 \wedge \dots \wedge A_k=a_k) = \frac{\Pr(C=c_i) \prod_{j=1}^k \Pr(A_j=a_j | C=c_i)}{\Pr(A_1=a_1 \wedge \dots \wedge A_k=a_k)} \quad (2)$$

对训练集来说,公式(2)中的概率容易给出近似的计算:

$$\hat{\Pr}(A_j=a_j | C=c_i) = \frac{\text{count}(A_j=a_j \wedge C=c_i)}{\text{count}(C=c_i)} \quad (3)$$

可见通过公式(3)对训练数据集求得相应的近似概率,就可以最终得出类的最大可能性的概率估计。该公式是算法的一个关键点,为了防止(3)式出现0的情况,当(3)式实际计算为0时,可以直接指定(3)式的结果为 $0.5/N$,这里 N 为例子的总数。

假定只有两类,将这两类分别称为0类和1类; a_1 到 a_k 表示一个给定测试集的属性; $b_0 = \Pr(C=0)$, $b_1 = \Pr(C=1) = 1 - b_0$; $p_{j0} = \Pr(A_j=a_j | C=0)$; $p_{j1} = \Pr(A_j=a_j | C=1)$ 。则:

$$p = \Pr(C=1 | A_1=a_1 \wedge \dots \wedge A_k=a_k) = (\prod_{j=1}^k p_{j1}) b_1 / z \quad (4)$$

$$q = \Pr(C=0 | A_1=a_1 \wedge \dots \wedge A_k=a_k) = (\prod_{j=1}^k p_{j0}) b_0 / z \quad (5)$$

这里 z 是一个固定常量,将上两式两边取对数后相减得:

$$\log p - \log q = (\sum_{j=1}^k \log p_{j1} - \log p_{j0}) + \log b_1 - \log b_0 \quad (6)$$

这时,取 $w_i = \log p_{i1} - \log p_{i2}$, $b = \log b_1 - \log b_0$, 则上式转化为:

$$\log(1-p)/p = -\sum_{i=1}^k w_i - b \quad (7)$$

在两边取指数,并且重新排列:

$$P = 1/[1 + e^{-\sum_{i=1}^k w_i - b}] \quad (8)$$

为计算该式,一般来说,假定属性 A_j 具有 $v(j)$ 个可能的属性值,那么当:

$$w_{jv} = \log \Pr(A_j = a_{jv} | C=1) - \log \Pr(A_j = a_{jv} | C=0) \quad (9)$$

$$1 \leq j \leq v(j)$$

时,上式转化为:

$$\Pr(C(x)=1) = \frac{1}{1 + e^{-\sum_{j=1}^J \sum_{v=1}^{v(j)} I(A_j=a_{jv}) w_{jv} - b}} \quad (10)$$

这里, I 是一个指示函数:如果 φ 为真,那么 $I(\varphi)=1$, 如果 φ 为假,那么 $I(\varphi)=0$.

在实际计算中上式只需根据公式(3)的思想即可得出计算结果。

公式(10)实际描述的是一个具有 sigmoid 型的感知器激活函数。该函数分别输入每一个属性 A_j 的 $v(j)$ 个可能的值。因此朴素贝叶斯分类器在特定输入场合上的表现能力等价于感知器。

进一步的研究表明,朴素贝叶斯分类器可以被推广到针对数值属性的逻辑回归方法。

在此考虑公式(3),如果 A_j 具有离散值,那么 $\text{count}(A_j=a_j \wedge C=c_i)$ 能够从训练例中直接计算出来。如果 A_j 是连续的数字值,那么就需要离散化该属性,使用非监督离散化将该属性离散化到 M 个等宽度的区间内,一般 $M=10$ 。也可以使用更复杂的离散化方法如监督离散化方法。

让每一个 A_j 是一个数字属性(整数值,或实数值),逻辑回归假设模型:

$$\log \frac{\Pr(C=1|A_1=a_1, \dots, A_k=a_k)}{\Pr(C=0|A_1=a_1, \dots, A_k=a_k)} = \sum_{j=1}^k b_j a_j + b_0 \quad (11)$$

经过公式(6)到(8)同样的处理过程后:

$$p = 1/[1 + e^{-\sum_{j=1}^k b_j a_j - b_0}] \quad (12)$$

显然,这是一个具有 sigmoid 激活函数,每个属性作为输入的单节点的感知器,并且每个属性值以它们的大小编码。用一个 $\varphi=a_j$ 的函数 $f(\varphi)$ 替换 $b_j a_j$, 如果 A_j 的范围被分割为 M 个部分,那么第 i 部分在 $[c_{j(i-1)}, c_{jp}]$, 则该函数就是:

$$b_j a_j = f_j(a_j) = \sum_{i=1}^M b_{ji} I(c_{j(i-1)} \leq a_j < c_{jp}) \quad (13)$$

这里每一个 b_{ji} 是一个常数。联合公式(12)(13)则推出公式(10)的一个新的版本:

$$\Pr(C(x)=1) = \frac{1}{1 + e^{-\sum_{j=1}^J \sum_{i=1}^M b_{ji} I(c_{j(i-1)} \leq a_j < c_{jp}) - b_0}} \quad (14)$$

这就是最终的回归函数。因此朴素贝叶斯分类是逻辑回归的一种非参数化,非线性的扩展。通过设置 $b_{ji} = (c_{j(i-1)} + c_{jp})/(2b_j)$ 能够近似地得到标准逻辑回归公式。

2 增强型方法

增强型方法的总的思想是学习一系列分类器,在这个序列中每一个分类器对它前一个分类器导致的错误分类给予更大的重视,尤其是:在学习完分类器 H_k 之后,增强型增加了由 H_k 导致分类错误的训练例的权值,并且通过重新对训练例计算权值,再学习下一个训练例 H_{k+1} 。这个过程重复 T 趟。最终的分类型从这一系列分类器中得出。

在这个过程中,每个训练例被赋予一个相对权值,如果一个训练例被分类器错误分类,那么就相应地增加该训练例的相对权值,使得在一次学习时,分类器对该训练例代表的情况更加重视。

通过增强可以把一个“弱”的算法增强为一个“强”的算法。

增强型方法中对两类分类问题的主要方法是: AdaBoost 算法,由 Freund 和 Schapire^[1]给出:

AdaBoost 算法

Input:

有 N 个训练例: $\langle (x_1, y_1), \dots, (x_N, y_N) \rangle$, $i=1, \dots, N$ 。

N 个例子上的分布 D : w_i 为每个训练例的权。

弱学习算法(在此为朴素贝叶斯)。

t 为趟数, T 为总的趟数。

Initialize

w_i 被初始化为 $w_i^{(1)} = 1/N$ 。

Do for $t=1, \dots, T$ do

(1) 给定权值 $w_i^{(t)}$ 得到一个假设 $H^{(t)}: X \rightarrow [0, 1]$ 。

(2) 得到 $H^{(t)}$ 的误差, 为 $\epsilon^{(t)} = \sum_{i=1}^N w_i^{(t)} |y_i - h_i^{(t)}(x_i)|$ 。

(3) 求出 $\beta^{(t)} = \epsilon^{(t)} / (1 - \epsilon^{(t)})$ 。

(4) 得到 $w_i^{(t+1)} = w_i^{(t)} (\beta^{(t)})^{1-y_i - h_i^{(t)}(x_i)}$ 。

(5) 正规化 $w_i^{(t+1)}$, 使其总和为 1.0。

End Do

Output:

$$(6) h_T(x) = \begin{cases} 1 & \text{若 } \sum_{t=1}^T \left(\log \frac{1}{\beta^{(t)}} \right) h^{(t)}(x) \geq \\ & \frac{1}{2} \sum_{t=1}^T \log \frac{1}{\beta^{(t)}} \\ 0 & \text{其它} \end{cases}$$

对算法的分析为: 在这里假设每一个分类器都是实际有用的, $\epsilon^{(t)} < 0.5$, 即在每次分类的结果中, 正确分类的结果始终大于错误分类的结果。那么可以得出 $\beta^{(t)} < 1$, 当 $|y_i - h_i^{(t)}(x_i)|$ 增加时, $w_i^{(t+1)}$ 增加。因此显然算法满足了增强的思想。

值得说明的是, 在对朴素贝叶斯进行增强时, 关于假设 $h^{(t)}(x)$:

1) $h^{(t)}(x)$ 采用公式(10)计算得出, 计算的结果或

者为0或者为1,但 w_i 参数的计算不同于原式。

2) 公式(9)中,关于 $Pr(a_i = a_{i,r} | C=c)$ 的计算:考虑公式(3),未增强的情况下, $count(condition)$ 计数操作的基准原来是1,例如有 k 个例子满足 $condition$,那么 $count(condition)$ 的值就是 k ,在增强的情况下 $count(condition)$ 计数操作的基准变为每一个例子的权值 w_i ,如果有 k 个例子满足 $condition$,那么现在 $count(condition)$ 的值就是这 k 个例子的每个例子的权值 w_i 的和。通过每个例子的权值,体现出增强的意义。

3) 所谓算法的输出是指:当给定一个新的 x ,通过算法的第(6)步,就可以利用以前的每次增强学习结果,通过投票输出最终的假设结果。

4) 正规化就是:

$$w_i^{(t+1)} = \frac{w_i^{(t)} e^{-\alpha f_i(x)}}{\sum_{i=1}^N w_i^{(t)} e^{-\alpha f_i(x)}}$$

算法最终的联合假设 H 可以给定为:

$$H(x) = \frac{1}{1 + \prod_{t=1}^T (\beta^{(t)})^{\sum_{i=1}^N w_i^{(t)} f_i(x)}}$$

其中:

$$r(x) = \frac{\sum_{i=1}^N (\log 1/\beta^{(i)}) H^{(i)}(x)}{\sum_{i=1}^N \log 1/\beta^{(i)}}$$

现在可以说明,被增强的朴素贝叶斯分类器表现上等价于一个具有单隐层的多层感知器。让: $\alpha = \prod_{t=1}^T \beta^{(t)}$, $v^{(i)} = \log \beta^{(i)} / \log \alpha$, 那么:

$$\begin{aligned} H(x) &= \frac{1}{1 + \alpha^{\sum_{i=1}^N v^{(i)} H^{(i)}(x)}} \\ &= \frac{1}{1 + e^{\sum_{i=1}^N 2k \log v^{(i)} H^{(i)}(x) - \log \alpha}} \\ &= \frac{1}{1 + e^{\sum_{i=1}^N 2 \log v^{(i)} H^{(i)}(x) - \sum_{i=1}^N k \log \beta^{(i)}}} \end{aligned}$$

即,单个分类器的输出乘以权值的和应用到一个 sigmoid 函数构成联合分类器的输出。既然单个朴素贝叶斯分类器等价于一个感知器,那么联合分类器等价于一个具有单隐层的感知器网络。

以上推导说明增强型朴素贝叶斯学习可以看作是反馈型神经网络一个替代。

当朴素贝叶斯用于多类分类时,相应的增强型方法为:

AdaBoost-M1 算法

Input:

有 N 个训练例: $\{(x_1, y_1), \dots, (x_N, y_N)\}$, $i=1, \dots, N$, 其标注为 $y_i \in Y = \{1, \dots, k\}$ 。

N 个例子上的分布 D : w_i 为每个训练例的权。

弱学习算法(在此为朴素贝叶斯)。

ϵ 为趟数, T 为总的趟数。

Initialize

w_i 被初始化为 $w_i^{(0)} = 1/N$ 。

Do for $t=1, \dots, T$ do

(1) 给定权值 $w_i^{(t)}$ 得到一个假设 $H^{(t)}: X \rightarrow Y$ 。

(2) 得到 $H^{(t)}$ 的误差, 为 $\epsilon^{(t)} = \sum_{i=1}^N w_i^{(t)} [y_i \neq h_i^{(t)}(x_i)]$ 。

(3) 求出 $\beta^{(t)} = \epsilon^{(t)} / (1 - \epsilon^{(t)})$ 。

(4) 得到 $w_i^{(t+1)} = w_i^{(t)} (\beta^{(t)})^{1 - \epsilon_i^{(t)}} = h_i^{(t)}(x_i)$ 。

(5) 正规化 $w_i^{(t+1)}$, 使其总和为 1.0。

End Do

Output:

$$(6) h_t(x) = \arg \max_{y \in Y} \sum_{i=1}^T \left(\log \frac{1}{\beta^{(i)}} \right) [h^{(i)}(x) = y]$$

这里的 $[\pi]$ 运算表示为: 如果 π 为真那么 $[\pi] = 1$, 如果 π 为假那么 $[\pi] = 0$ 。

3 计算复杂性讨论

假设例子有 f 个属性, 每一个属性有 v 个属性值。那么由公式(10)导出的一个朴素贝叶斯分类器有 $fv + 1$ 个参数。这些参数被累计学习 $2fv + 2$ 次。在每一次学习中, 每一个训练例的每一个属性值都会使得精度得以提高。因此, n 个训练例的时间复杂度就是 $O(nf)$, 与 v 无关, 这样的时间复杂度本质上就是最佳的。在增强情况下, 每一趟的时间复杂度也是 $O(nf)$, 因此 T 趟的话, 就是 $O(Tnf)$, 因为 T 是一个常数, 所以时间复杂度仍然是 $O(nf)$ 。

在生成朴素贝叶斯分类器时, 其主要运算是计数运算(参见公式(3))。训练例可以从磁盘或磁带上顺序地进行处理。内存的需求是 $O(fv)$, 与训练例无关。在增强情况下, 内存的需求是 $O(Tfv)$ 。因此该方法非常适合在大数据库中进行知识发现。数据集不需装入内存, 仍旧可以保存在磁盘或磁带上。

4 实验结果

以“德国信贷”数据集为例, 这个数据集有 700 个正例和 300 个反例, 17 个离散属性, 3 个连续属性。文献可查的最好的预测结果是: 错误率为 25.3%, 使用了一个朴素贝叶斯分类器和 5 重交叉验证(five-fold cross-validation)^[10]。

下表显示了在该数据集上增强型朴素贝叶斯分类结果, 从表上可以看出当联合 3 个朴素贝叶斯分类器时, 有最好的结果。用增强改进朴素贝叶斯方法缩减错误率只需要很少的趟数。在所有的数据集上, 做相同的实验都得出了相似的结果。

趟数	1	2	3	4	5	6	7	8	9	10	...	100
错误率	25.124	324.024	424.724	925.225	125.325	425.4	...	25.7				

考虑到该方法运行速度、对内存的需求、趟数, 可以得出, 在硬件条件允许下, 增强型朴素贝叶斯方法可以被看作是一种在任何时候都可以运行的学习算法。

然而, 增强方法应用到朴素贝叶斯存在如下问题^[4]:

1) 从增强的思想来看, 当训练集中存在噪音时, 增

强会把这些噪音当作有用的信息通过权值而放大,这样会降低增强的性能,当噪音很多时,甚至增强会导致更糟的结果,所以当训练集中存在噪音时,增强不能很好地处理这些噪音。

2)虽然理论上增强方法能够确保训练集的差错率为0,但当其应用于朴素贝叶斯时,却不能保证是这样,训练集的差错始终存在。

参考文献

- 1 Duda R, Hart P. Pattern Classification and Scene Analysis. John Wiley and Sons, New York, 1973
- 2 Freund Y, Schapire R E. A decision-theoretic generalization of on-line learning and an application to boosting. In: Proc. of the Second European Conf. on Computational Learning Theory, 1995. 23~37
- 3 Elkan C. Boosting and naive Bayesian learning. [Technical Report No CS97-557]. Department of Computer Science and Engineering, University of California, September 1997
- 4 Bauer E, Kohavi R. An Empirical Comparison of Voting Classification Algorithms: Bagging, Boosting, and Variants.
- 5 Bechara A, et al. Deciding advantageously before knowing the advantageous strategy. Science, 1997, 275: 1293~1295
- 6 Blum A L, Rivest R L. Training a 3-node neural network is NP-complete. Neural Networks, 1992, 5(1): 117~127
- 7 Bures J, et al. Place cells and place navigation. Proc. of the National Academy of Sciences, 1997, 94(1): 343~350
- 8 Domingos P, Pazzani M. Beyond independence: Conditions for the optimality of the simple bayesian classifier. In: Proc. of the 13th Intl. Conf. on Machine Learning. Morgan Kaufmann Publishers, Inc., 1996. 105~112
- 9 Dougherty J R, et al. Supervised and unsupervised discretization of continuous features. In: Proc. of the 12th Intl. Conf. on Machine Learning. Morgan Kaufmann Publishers, Inc., 1995. 194~202
- 10 Friedman N, Goldszmidt M. Building classifiers using Bayesian networks. In: Proc. of the National Conf. on Artificial Intelligence. Portland, Oregon, AAAI Press, August 1996
- 11 Garner W R, Suthoff D. The effect of goodness on encoding time in visual pattern discrimination. Perception and Psychophysics, 1974, 16(3): 426~430
- 12 Kalocsai P, et al. Predicting face recognition from a spatial filter similarity space. Personal communication, 1996
- 13 Kruschke J K. Human category learning. Implications for backpropagation models. Connection Science, 1993, 5(1): 3~36
- 14 Lehky S R, Sejnowski T J. Network model of shape-from-shading: neural function arises from both receptive and projective fields. Nature, 1988, 333(6172): 452~454
- 15 Mezard M, Nadal J P. Learning in feedforward layered networks: The tiling algorithm. J. of Physics A, 1989, 22(12): 2191~2203
- 16 Minsky M, Papert S. Perceptrons, an introduction to computational geometry. MIT Press, 1969
- 17 Quinlan J R. Boosting, bagging, and C4.5. In: Proc. of the National Conf. on Artificial Intelligence. Portland, Oregon, AAAI Press, 1996
- 18 Ripley B. Pattern Recognition and Neural Networks. Cambridge University Press, 1996
- 19 Schapire R, et al. Boosting the margin: a new explanation for the effectiveness of voting methods. In: Proc. of the 14th Intl. Conf. on Machine Learning. Morgan Kaufmann Publishers, Inc., 1997. To appear

(上接第92页)

业的计算活动,进而整个营运处理集成起来,形成一个企业级的集成框架。

目前基于 Agent 进行 workflow 系统管理尚处于摸索阶段,许多地方有待于完善。规划和协商算法有待进一步的研究和开发,另外还有 workflow 的安全问题,以及加入移动 Agent 技术解决分布式计算环境中更复杂、灵活的问题。

参考文献

- 1 The Workflow Management Coalition: Workflow Facility Specification, Document Number WfMC-TC-2101, Draft D. 6, 8-March-97
- 2 The Workflow Management Coalition: Workflow Standard-Interoperability Abstract Specification, Document Number WfMC-TC-1012, 20 October 1996, Version 1.0
- 3 高济. 基于表示本体论的智能系统开发. 计算机研究与发展, 1996, 33(11): 801~807
- 4 Alty J L, et al. ADEPT-Advanced Decision Environment for Process Tasks: Overview & Architecture. In: Proc. BCS Expert Systems 94 Conf. (Applications Track), Cambridge, UK, 1994, 359~371
- 5 Casati F, et al. WIDE Workflow Model and Architecture. [CTIT Technical Report 96-19], 1996. 4
- 6 Mohan C, et al. Exotica: A Research Perspective on Workflow Management Systems. Data Engineering Bulletin, 1995, 18(1)