文章编号:1001 - 9081(2010)06 - 1530 - 03

基于 Hellinger 距离的特征选择算法

李伟漳[1,2],贾修一[3]
(1. 南京航空航天大学 高新技术研究院,南京 210016;
2. 南京航空航天大学 信息科学与技术学院,南京 210016; 3. 南京大学 计算机科学与技术系,南京 210093)
(amy. vivilee@ gmail. com)

摘 要:针对数据挖掘中的特征选择问题,依据 Hellinger 距离的特性,研究了两种 Hellinger 距离的定义方式,提出了基于 Hellinger 距离的特征选择方法,设计了两种相应的算法。不同数据集上的实验结果表明了新算法选择的特征的有效性。与其他特征选择算法的对比可发现:这两种算法选择的特征个数少且对 C4.5 分类精度较好。

关键词:特征选择; Hellinger 距离; 数据挖掘
中图分类号: TP18 **文献标志码**:A

Feature selection algorithm based on Hellinger distance

LI Wei-wei[1,2], JIA Xiu-yi[3]
(1. *Academy of Frontier Science, Nanjing University of Aeronautics and Astronautics, Nanjing Jiangsu* 210016, *China*;
2. *College of Information Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing Jiangsu* 210016, *China*;
3. *Department of Computer Science and Technology, Nanjing University, Nanjing Jiangsu* 210093, *China*)

Abstract: To solve the feature selection problem, two kinds of definitions of Hellinger distance were studied in this paper, and the corresponding feature selection algorithms based on Hellinger distance were also proposed. The experiments on different data sets show the efficiency of the two algorithms. Compared with other feature selection algorithms, the feature selection algorithms based on Hellinger distance can get fewer features, which are useful for C4.5 and can improve the average accuracy of the classification in the learned data sets.

Key words: feature selection; Hellinger distance; data mining

0 引言

在数据挖掘过程中,针对实际应用中的一些高维问题,很多学者对特征选择(Feature Selection)进行了广泛的研究。对于特征选择方法主要分为 Filter 模型和 Wrapper 模型[1],其中 Filter 模型主要对特征通过一些度量进行排序,进而选择合适的特征;而 Wrapper 方法则通过一个学习器(通常是分类器)对所选择的特征集合进行评价,进而选择合适的特征。在 Filter 模型中评价特征的相关性或重要性是通过特定的度量所决定的,不同的度量会产生不同的特征选择算法,如基于各种距离度量的特征选择算法、基于信息熵理论的特征选择算法等[1]。

基于距离的特征选择算法是一类典型的特征选择方法,由于通过各种距离能够度量特征之间的相关性和重要性,很多学者进行了相应的研究,提出了如基于 Bhattacharyya 距离[2]、Hausdorff 距离[3]、I-divergence 和 Vasershtein 距离等[4]特征选择方法。Hellinger 距离作为一种距离度量,Cieslak 等人[5]基于其设计了相应的决策树算法,并分析和实验验证了该距离度量作为选择合适特征进行建树的有效性,特别是在不平衡数据上;Lee 等人[6]也将其应用在最近邻分类器上。本文在此基础上针对 Hellinger 距离的特性,设计了两种特征选择算法。实验表明,这两种算法能够得到较小的特征集合,且选出的特征对 C4.5 具有更好的分类效果。

1 Hellinger 距离及其应用

1.1 Hellinger 距离定义

Hellinger 距离是一种能够体现两个分布之间距离的度量[5,7]。假设在可度量空间(Θ,λ)中,P 和 Q 分别代表对应参数 λ 的两个连续分布,则这两个分布之间的 Hellinger 距离定义为:

$$d_H(P,Q) = \sqrt{\int_\Omega (\sqrt{P} - \sqrt{Q})^2 d\lambda}$$

等价于:

$$d_H(P,Q) = \sqrt{2\left(1 - \int_\Omega \sqrt{PQ} d\lambda\right)}$$

对于可数空间Φ,也就是当分布是离散型的,两个分布之间的 Hellinger 距离可定义为:

$$d_H(P,Q) = \sqrt{\sum_{\varphi\in\Phi} \left(\sqrt{P(\varphi)} - \sqrt{Q(\varphi)}\right)^2}$$

由公式可以得出 Hellinger 距离具有如下性质:

$$\begin{cases} 0 \leqslant d_H(P,Q) \leqslant \sqrt{2} \\ d_H(P,Q) = d_H(Q,P) \end{cases}$$

两个分布之间的 d_H 是非负和对称的。在本文中主要采用可数空间里分布之间的 Hellinger 距离定义形式。

1.2 Hellinger 距离的应用

在数据挖掘应用中,Cieslak 和 Chawla 针对不平衡数据提出了基于 Hellinger 距离度量的决策树算法。针对两类问题,

收稿日期:2010 - 01 - 04;**修回日期**:2010 - 03 - 10。 **基金项目**:江苏省自然科学基金资助项目(BK2009233)。

作者简介:李伟漳(1981 -),女,湖北宜昌人,博士研究生,主要研究方向:数据挖掘; 贾修一(1983 -),男,山东日照人,博士研究生,CCF 会员,主要研究方向:粗糙集理论、机器学习、自然语言处理。

X_+ 为正类,X_- 为负类,在可数空间内,假设特征 f 依据其取值被划分为 p 类,连续型数据先离散化处理,则两类在特征 f 上的分布之间的 Hellinger 距离定义为:

$$d_{\mathrm{H}}(X_+,X_-)=\sqrt{\sum_j^p\left(\sqrt{\frac{|X_{+j}|}{|X_+|}}-\sqrt{\frac{|X_{-j}|}{|X_-|}}\right)^2}$$

若用概率 $P(X_j\mid X_+)$ 表示对象在划分为正类的情况下其特征 f 上的取值为 X_j 的条件概率,则公式可表示为:

$$d_{\mathrm{H}}(X_+,X_-)=\sqrt{\sum_j^p\left(\sqrt{P(X_j\mid X_+)}-\sqrt{P(X_j\mid X_-)}\right)^2}$$

在文献[5]中分析到,由于该距离公式并不考虑分布的先验概率,所以该距离对于不平衡数据不是很敏感。针对不平衡数据,基于该度量的决策树算法会取得较好的效果。

Lee 等人则通过信息论角度来应用 Hellinger 距离到分类器中,基于 Hellinger 距离来计算特征 f 对决策特征贡献的信息量。对于目标特征(通常是指决策特征)T,t_i 表示 T 的一个取值,f_j 为特征 f 的一个取值,令 $P(t_i)$ 和 $P(t_i\mid f_j)$ 分别表示 T 的一个先验概率和在特征 f 取值为 f_j 时的条件概率,则特征 f 取值为 f_j 时给予目标特征 T 的信息量定义为:

$$H(T\mid f=f_j)=\sqrt{\sum_i\left(\sqrt{P(t_i)}-\sqrt{P(t_i\mid f=f_j)}\right)^2}$$

特征 f 对目标特征 T 的信息量定义为:

$$H(T\mid f)=\sum_{f_j}P(f_j)H(T\mid f=f_j)$$

特征集合中每个特征的权值定义为:

$$\omega_T(f)=\frac{H(T\mid f)}{\sum_f H(T\mid f)}$$

比较这两种定义方式,Cieslak 等人定义的 Hellinger 距离体现了两类在同一特征上的距离,取值越大,说明基于该特征区分这两类能力越强,以此作为决策树的选择标准。而 Lee 等人定义的基于具体特征取值的 Hellinger 距离体现了该特征值对决策类分布的区分能力,若取值越大,则说明使用该特征值更易区分各个决策类。

通过对这两种 Hellinger 距离使用的研究可以看出,定义的这两种距离都能体现特征对最终分类的影响程度。而当我们考虑整个特征空间,将整个条件特征(相对于目标特征或决策特征而言)作为一个特征时,该特征对最终分类的影响程度是最大的,也就是说不考虑实际应用中的噪声问题,只就数据提供的信息量来说,特征越多,信息量就越多,则特征选择的目标就是选取合适的子集使其对最终分类提供相同或稍少于所有特征集合的信息量。对于特征子集与基于该子集上的这两种 Hellinger 距离都是单调的,限于篇幅所限,在此就不予证明。基于此我们设计了两种基于 Hellinger 距离的特征选择算法。

2　基于 Hellinger 距离的特征选择算法

综上可知,选取特征子集的准则就是使得基于选取的子集的 Hellinger 距离值要和基于所有特征集合的 Hellinger 距离值相同或者低于某个特定的阈值。假设这个特定的阈值为 α,该阈值表示选取子集后的 Hellinger 距离的值和基于所有特征集合的 Hellinger 距离的值的比值大于等于 $\alpha(0\leqslant\alpha\leqslant1)$。针对这两种定义,本文设计了两种特征选择算法。

考虑到所有特征子集的可能组合是个 NP 问题,所以采用启发式算法,先求出所有单个特征的 Hellinger 距离值,依据大小对其排序,得到一个特征的序列集。先选取具有最大值的特征,并将其从序列集中移出,判断其值与包含所有特征集合的 Hellinger 距离值是否满足大于等于 α 的关系,若满足则将该特征加入到结果集中结束程序;若不满足,判断增加的该特征是否对已有的特征集合有用,也就是使基于该集合的 Hellinger 距离值增大,若值不发生变化,则认为该特征对已有的特征集合是冗余的,删除该特征,再挑选剩下序列集中取值最大的特征加入到结果集中,重新计算新得到的结果集的 Hellinger 距离值,直到得到满足条件的特征集合为止。

算法主框架描述如下。

输入:训练集 M;特征集 $At=C\cup D$,条件属性集 C,决策属性集 D;阈值 α。

输出:特征子集 $result$。

```
1)  result = ∅;
2)  temp = ∅;
3)  H_C = CalulateHellingerDistance(M,C);
        // 计算基于整个条件特征集合的 Hellinger 距离值
4)  FOR each feature c_i in condition feature set C
5)    H_Ci = CalculateHellingerDistance(M, c_i);
                    // 计算每个特征的 Hellinger 距离值
6)  END FOR
7)  RH = Rank(H_Ci);                    // 从大到小排列
8)  m = max(RH);                        // 每次选取值最大的特征
9)  RH = RH - m;                        // 将其移出
10) temp += m;
11) H_last = H_temp = CalculateHellingerDistance(M, m);
12) IF H_temp/H_C ≥ α THEN
13)   result = m;                       // 得到结果并返回
14)   RETURN result;
15) ELSE
16)   WHILE RH is not empty
17)     m = max(RH);
18)     RH = RH - m;                    // 将其移出
19)     temp += m;
20)     H_temp = CalculateHellingerDistance(M, temp);
21)     IF H_temp/H_C ≥ α THEN
22)       result = temp;                // 得到结果并返回
23)       RETURN result;
24)     END IF
25)     IF H_temp == H_last THEN
26)       temp = temp - m;
          // 如果增加的特征并没有使得距离值发生改
          // 变,则移出
27)     END IF
28)   END WHILE
29) END IF
```

以上给出了特征选择的算法主框架,其核心部分在于函数 CalculateHellingerDistance 的计算。对于 Hellinger 距离的计算,设计了两种计算方式。

第一种计算特征的 Hellinger 距离函数定义如下。

FUNCTION1 CalculateHellingerDistance

输入:训练集 M,特征集 f。

输出:Hellinger 距离的值。

```
1)  hellinger = 0;n_D 为训练集 M 决策类个数
2)  FOR each distinct class pair of a, b in M
3)    FOR each value v of condition feature set f
4)      h += (sqrt(|M_v,a|/|M_a|) - sqrt(|M_v,b|/|M_b|))^2;
5)    END FOR
```

6) $hellinger += \sqrt{h}$;
7) END FOR
8) RETURN $hellinger/(n_D(n_D-1)/2)$;

第二种计算 Hellinger 距离方式如下。

FUNCTION2 CalculateHellingerDistance

输入：训练集 M，特征集 f。

输出：Hellinger 距离的值。

1) $hellinger = 0; n = size$ of training set M
2) FOR each value f_j of condition feature set f
3) FOR each value d of decision feature set D
4) $$h += \left(\sqrt{\frac{|M_d|}{n}} - \sqrt{\frac{|M_{f_j,d}|}{|M_{f_j}|}}\right)^2;$$
5) END FOR
6) $$hellinger += \frac{|M_{f_j}|}{n}\sqrt{h};$$
7) END FOR
8) RETURN $hellinger$;

通过这两种定义方式可以看出，第一种定义方式对应于 Cieslak 等人定义的 Hellinger 距离，体现了两类在同一特征之间的距离；第二种定义方式对应于 Lee 等人定义的 Hellinger 距离，体现了特征值对决策分布的影响程度。

3 实验

以 Weka 3.5.7[8] 为平台实现了本文设计的两个算法，并与 Weka 中的基于 GainRatio 和 χ^2 距离（ChiSquare）的两个特征选择算法进行比较。在选择出特征后，用 C4.5 作为分类器对数据进行分类，对比选择出来的特征对最终分类精度的影响。分类精确度测试方法采用 10 倍交叉验证。为方便表示，第一种计算 Hellinger 距离函数的算法简写为 HD1，第二种简写为 HD2，基于 GainRatio 的特征选择算法简写为 GR，基于 χ^2 距离的简写为 CS。

数据集选取的原则是尽量选择特征较多，不存在缺失值且适用于测试分类的数据集，本文选取了 11 个数据集，其中 German. Numer 和 Splice 来自文献[9]，其他均来自 UCI 数据集[10]，这 11 个数据集的详细情况描述如表 1 所示。

表 1 数据集详细描述

数据集	对象数	特征数	类别数
Arcene	100	10 000	2
Breast Cancer_wdpc	569	31	2
Breast Cancer_wdpc	198	33	2
Flags	194	29	6
German. Numer	1 000	24	2
Hill-Valley	606	100	2
ionosphere	351	34	2
Statlog(Landsat Satellite)	4 435	36	6
SPECT Heart	267	22	2
SPECTF Heart	267	44	2
Splice	2 175	60	2

关于各个算法的参数设置问题，对于算法 GR 和 CS 都是采用常用的 Ranker 的方法，其中选取特征的阈值都设置为 0，由于选择的数据特征都较多，而且由于 Hellinger 距离值的定义方式导致基于每个特征的距离值都较小，所以在 HD1 和 HD2 算法中，α 都设置为较小的0.001。通过实验表明，α 值提高一个数量级时，选择的特征非常少；降低一个数量级时选择的特征和原特征空间几乎相同，不具代表性。限于篇幅所限，在此就不予赘述。实验也从两方面进行了比较：一是选取的特征个数，二是依据选取的特征对分类效果的影响。实验结果如表 2～3 所示。

表 2 各个算法在数据集上选择的特征数

数据集	算法			
	GR	CS	HD1	HD2
Arcene	1 534	1 534	29	1
Breast Cancer_wdpc	27	27	2	2
Breast Cancer_wdpc	2	2	2	1
Flags	10	10	10	7
German. Numer	15	15	9	6
Hill-Valley	0	0	1	1
ionosphere	33	33	10	3
Statlog (Landsat Satellite)	36	36	8	8
SPECT Heart	22	22	21	21
SPECTF Heart	29	29	4	3
Splice	40	40	43	12

表 3 基于各算法选择的特征的 C4.5 分类精度 %

数据集	算法			
	GR	CS	HD1	HD2
Arcene	78.00	78.00	70.00	54.00
Breast Cancer_wdpc	92.80	93.50	93.00	90.50
Breast Cancer_wdpc	73.70	73.70	76.30	76.30
Flags	69.10	68.60	70.60	70.60
German. Numer	72.20	72.20	73.90	72.60
Hill-Valley	Null	Null	50.30	50.30
ionosphere	90.90	91.20	82.60	79.50
Statlog (Landsat Satellite)	86.20	86.20	87.00	86.40
SPECT Heart	70.40	70.40	70.40	70.40
SPECTF Heart	76.40	76.80	77.90	79.80
Splice	95.00	94.80	94.90	95.30
Average Rank	2.18	2.27	1.73	2.00

通过表 2 可以看出，HD1 和 HD2 两个特征选择算法所选取的特征要少于 GR 和 CS 特征选择算法；而且在使用较少特征情况下，对 C4.5 分类器影响性能并没有多大影响。相反，如表 3 所示，在 11 个训练集上得到的平均结果都要好于 GR 和 CS 所选择的特征，其平均排名都要高于 GR 和 CS。

4 结语

本文提出了基于 Hellinger 距离的特征选择算法，对于特征选择问题，基于两种 Hellinger 距离的定义方式，设计了两种对应的特征选择算法。实验结果表明在数据集上这两种特征选择算法都能取得较少的特征，而且基于选择的特征用 C4.5 进行分类，效果比较好，反映了算法的有效性。

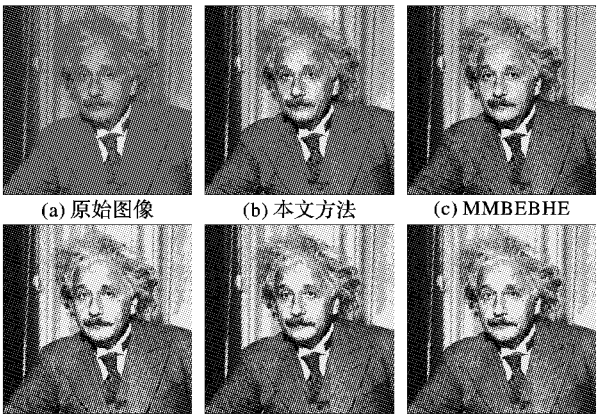
在文献[5]中分析到其设计的 Hellinger 距离不考虑决策类的先验分布，从而对不平衡数据更有效，本文没有对此进行实验上的验证，这将是我们下一步的研究工作。

参考文献：

[1] GUYON I, ELISSEEFF A. An introduction to variable and feature selection [J]. Journal of Machine Learning Research, 2003, 3: 1157－1182.

（下转第 1634 页）

用各种方法分别处理大小不同的图像，图像越大，相应处理时间就越长，以 256×256 的图 2 为原始图像从计算复杂度上对各种算法进行比较：



(a) 原始图像 (b) 本文方法 (c) MMBEBHE

(d) BBHE (e) DSIHE (f) HE

图 2 man 图像

HE 和 BBHE 最简单，运行时间也最快。HE 只需要进行一次直方图均衡，在 Matlab 环境下，时间为 0.204 000 s。BBHE 以原图像亮度均值为阈值对原直方图进行划分，需要进行两次均衡运算，所需时间为 0.313 000 s。而 DSIHE、MMBEBHE 和本文方法均需要在灰度级范围内做一个循环，以每个灰度值为分割阈值，对原直方图进行双直方图均衡，再根据处理结果选择最合适的分割阈值。即对于8 位灰度图像，最大需要做一个循环次数为 256 的循环，所以运行时间较长，这 3 种方法分别用时 8.078 000 s、8.094 000 s 和8.922 000 s。本文方法与 DSIHE 和 MMBEBHE 的用时时间差为 0.844 s 和 0.828 s，多的时间用于直方图的均匀化。

表 1 处理后的亮度均值与熵值对比

图像	处理方法	亮度均值	熵
图 1	本文方法	34.596	4.011 7
	MMBEBHE	44.857	4.000 8
	BBHE	43.555	3.992 4
	DSIHE	75.130	4.011 9
	HE	131.750	3.911 9
图 2	本文方法	107.920	4.661 9
	MMBEBHE	107.700	4.666 4
	BBHE	119.960	4.661 5
	DSIHE	114.970	4.668 8
	HE	119.700	4.661 1

综上所述，虽然 DSIHE、MMBEBHE 和本文算法的复杂度较高，用时较长，但处理效果较好，综合比较本文算法的处理效果最好。

4 结语

本文在双直方图均衡的基础上，采用了一种新的分割阈值选取方法，该阈值的选取综合考虑了图像的熵的亮度均值，并对均衡后的图像进行均匀化处理消除过增强现象。改进方法在对比度增强和保持图像细节上均好于原有双直方图均衡算法。

参考文献：

[1] GONZALEZ R C, WOODS R E. 数字图像处理[M]. 阮秋琦, 阮宇智, 译. 2 版. 北京: 电子工业出版社, 2003.

[2] PIZER S M, AMBURN E P, AUSTIN J D, *et al*. Adaptive histogram equalization and its variations [J]. Computer Vision Graphics, and Image Processing, 1987, 39(3): 355 – 368.

[3] KIM Y-T. Contrast enhancement using brightness preserving bi-histogram equalization [J]. IEEE Transactions on Consumer Electronics, 1997, 43(1): 1 – 8.

[4] WANG YU, CHEN QIAN, ZHANG BAEOMIN. Image enhancement based on equal area dualistic sub-image histogram equalization method [J]. IEEE Transactions on Consumer Electronics, 1999, 45(1): 68 – 75.

[5] CHEN S-D, RAMLI A R. Minimum mean brightness error bi-histogram equalization in contrast enhancement [J]. IEEE Transactions on Consumer Electronics, 2003, 49(4): 1310 – 1319.

[6] CHEN S-D, RAMLI A R. Contrast enhancement using recursive mean-separate histogram equalization for scalable brightness preservation [J]. IEEE Transactions on Consumer Electronics, 2003, 49(4): 1301 – 1309.

[7] 江巨浪, 张佑生, 薛峰, 等. 保持图像亮度的局部直方图均衡算法[J]. 电子学报, 2006, 34(5): 861 – 866.

[8] WANG CHAO, YE ZHONGFU. Brightness preserving histogram equalization with maximum entropy: A variational perspective [J]. IEEE Transactions on Consumer Electronics, 2005, 51(4): 1326 – 1334

[9] 陈钱, 柏连发, 张保民. 红外图像直方图双向均衡技术研究[J]. 红外与毫米波学报, 2003, 22(6): 428 – 430

[10] KIM J-Y, KIM L-S, HWANG S-H. An advanced contrast enhancement using partially overlapped sub-block histogram equalization [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2001, 11(4): 475 – 484.

(上接第 1532 页)

[2] XUAN G R, CHAI P Q, WU M H. Bhattacharyya distance feature selection [C]// Proceedings of the 13th International Conference on Pattern Recognition. Washington, DC: IEEE Computer Society, 1996, 2: 195 – 199.

[3] PIRAMUTHU S. The Hausdorff distance measure for feature selection in learning applications [C]// Proceedings of the 32nd Hawaii International Conference on System Sciences. Washington, DC: IEEE Computer Society, 1999.

[4] PAPANTONI-KAZAKOS P. Some distance measures and their use in feature selection, #7611 [R]. Houston: Rice University, Electrical Engineering Department, 1976.

[5] CIESLAK D A, CHAWLA N V. Learning decision tree for unbalanced data [C]// ECML/PKDD. Berlin: Springer-Verlag, 2008, 1: 241 – 256.

[6] LEE C H, SHIN D G. Using Hellinger distance in a nearest neighbor classifier for relational databases [J]. Knowledge-Based Systems, 1999, 12(7): 363 – 370.

[7] RAO C. A review of canonical coordinates and an alternative to correspondence analysis using Hellinger distance [J]. Questiio, 1995, 19(1/3): 23 – 63.

[8] Weka 3: Data mining software in Java [EB/OL]. [2009 – 12 – 20]. http://www.cs.waikato.ac.nz/ml/weka/.

[9] CHANG C. LIBSVM: A library for support vector machines [EB/OL]. [2009 – 12 – 20]. http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/.

[10] UCI: Machine learning repository [EB/OL]. [2009 – 12 – 20]. http://archive.ics.uci.edu/ml/.