

文章编号:1001-9081(2008)07-1692-04

一种基于正交法和扩展卡方检验的关联分类算法

孙 蕾,李军怀

(西安理工大学 计算机科学与工程学院, 西安 710048)

(darcy007007@163.com)

摘 要:针对几种典型分类算法中存在的诸如分类器性能较低和算法效率不高等问题,提出了一种基于正交法和扩展 χ^2 检验的分类算法 ERAC。算法首先通过正交法产生所有的频繁项集和关联规则,然后采用一种扩展 χ^2 检验来对规则进行分级和修剪,有效减少分类器的规则数目。试验结果表明,该算法与 CBA 等算法相比较具有较高的分类准确率和运行效率。

关键词:分类算法; 关联规则; 正交法; 卡方检验

中图分类号: TP311.1 **文献标志码:** A

Multi-class associative classification based on intersection method and extended chi-square testing

SUN Lei, LI Jun-huai

(College of Computer Science and Engineering, Xi'an University of Technology, Xi'an Shaanxi 710048, China)

Abstract: Given that there exist some defects in several typical classification algorithms, such as poor classification performance and long running time when the algorithm efficiency is not high, this paper proposed a classification algorithm ERAC based on intersection method and extended chi-square testing. This algorithm first produced, through intersection method, all the frequency items and association rules. Then it conducted classification and pruning on rules using an extended testing method, reducing the number of rules for classification effectively. Subsequent experiments prove that new method, compared with the CBA algorithm, has higher classification accuracy and operating efficiency.

Key words: classification algorithm; association rules; intersection method; chi-square testing

0 引言

构建精确而有效的分类器是数据挖掘中的一个重要任务。关联分类技术一般包括两个步骤:一是发现所有规则右部(RHS)为类标签的分类关联规则(class association, CARs)^[1];二是从已发现的 CARs 中选择优先级高的规则来覆盖训练集。规则的优先级往往根据分类关联规则的置信度、支持度、规则长度或一般分类规则质量标准进行评价^[2]。

目前,典型的分类算法有基于分类关联规则的分类算法(CBA)^[3],基于多个分类关联规则的分类算法(CMAR)^[4],一个懒的(Lazy)分类关联规则剪枝方法^[5],预测型关联规则的分类算法(CPAR)^[6],基于暗示强度(Implicitly Strength) CBA 改进算法^[7],通过利用组合算法中推进原理的关联分类算法^[8]。

以上介绍的关联分类算法的基本思想是利用现有关联规则挖掘算法产生所有频繁项目,并使用这些频繁项目集构造分类器,它们同时考虑所有属性,取得了比同样基于规则的决策树分类算法 C4.5 更好的分类效果。但是,此类算法仍然存在一些不足:1)算法的研究多集中在规则筛选的研究,对于规则怎样排序能提高分类器性能还有待进一步的研究和发展;2)它们把训练集作为一个整体,当训练集非常大时,挖掘关联规则的效率很低;3)算法挖掘频繁项集的步骤很多,当频繁项集的数量非常庞大时,会消耗大量的运行时间,算法效率

不高。

本文提出了一种基于正交法和扩展 χ^2 检验的 ERAC (Effective Rules based on Associative Classification) 算法,对经典算法的几个方面不足进行了改进。首先,通过正交法只需一步即可产生所有的频繁项集和关联规则,该方法不像其他方法需要分几步来产生频繁项集和关联规则,在运行时间和存储空间方面的效率高于原有的其他方法;然后,采用一种扩展 χ^2 检验来对规则进行分级和修剪,从而有效减少分类器的规则数目,最终达到提高分类准确率和运行效率的目的。

1 问题描述及相关定义

基于关联规则的多类别分类是针对多类别数据挖掘频繁模式、建立分类器,实现对未知样本分类的方法。可以分为三个步骤:挖掘频繁模式;建立分类器;对测试集进行预测。

给出相关定义如下:给定数据集 D , 令 D 有 n 个属性和 $|D|$ 行的训练数据集 $A_1, A_2, A_3, \dots, A_n$, I 为 D 中所有项的集合, C 为样本类别, Y 为分类关联规则, A 为项的集合(项集), S 为 Y 的支持数, sup 为 Y 的支持度, $conf$ 为 Y 的置信度。

定义 1 Y 具有如下形式: $Y: A \Rightarrow C$ 。

定义 2 Y 的支持数 S 等于 D 中包含 A 且类别为 C 的样本个数。

定义 3 一个规则 Y 在 D 中实际出现次数 $N(r)$ 是 D 中与 Y 的条件匹配。

收稿日期:08-01-07;修回日期:2008-04-24。

基金项目:国家 863 项目(2007AA010305);陕西省自然科学基金资助项目(2005F05)。

作者简介:孙蕾(1982-),女,河北怀安人,硕士,主要研究方向:数据挖掘;李军怀(1970-),男,陕西宝鸡人,副教授,博士,主要研究方向:服务计算、数据挖掘。

定义4 Y 的支持度定义为: $\text{sup}(Y) = \frac{S}{|D|}$ 。

定义5 Y 的置信度定义为: $\text{conf}(Y) = \frac{S}{N(r)}$ 。

定义6 规则库中一个规则的最小支持度表示为 Minsup 。

定义7 规则库中一个规则的最小置信度表示为 Minconf 。

2 ERAC 算法设计与实现

2.1 正交法

为了挖掘频繁项集和关联规则生成的效率,本文基于文献[9]提出了一种改进的正交法。使用正交法只需要对训练数据集扫描一次来计算单一项出现的次数,从而确定出超过 Minsup 阈值的项集,并按照其在排列中的位置(行号)来储存这些项。然后,通过对已发现的频繁项的行号取交集,可以很容易的得到剩余的涉及一个以上属性的频繁项。再对频繁单一项使用行号,来得到涉及一个以上的项以及它的规则的支持度和置信度阈值。

由于只需要对训练数据集扫描一次来发现和生成规则,该方法在运行时间和存储空间方面的效率是很高的,因为它不依赖于传统的需要多次扫描的方法。正交法的算法流程如下:

输入:训练数据(D),支持度和置信度阈值(MinSup and MinConf)

输出:用于分类的关联规则

- 1) Scan D for the set S of frequent single items
- 2) Do
- 3) For each pair of disjoint items I_1, I_2 in S
- 4) If $\langle I_1 \cup I_2 \rangle$ passes the MinSup threshold
- 5) $S \leftarrow S \cup \langle I_1 \cup I_2 \rangle$
- 6) Until no items which pass MinSup are found
- 7) For each item I in S
- 8) Generate all rules $I \rightarrow c$ which pass the MinConf threshold
- 9) Rank all rules generated
- 10) Remove all rules $I' \rightarrow c'$ from S where
- 11) there is some rule $I \rightarrow c$ of a higher rank and $I \subseteq I'$

2.2 扩展 χ^2 检验

挖掘频繁项集生成的关联规则数目可能变得非常巨大,为了使分类器能高效地对数据集分类,需要修剪规则以除去冗余的干扰信息。对关联规则进行分级是很重要的。CBA和CMAR主要按置信度等级来给规则分级,方法如下:给出两个规则 R_1 和 R_2 ,如果 R_1 被认为比 R_2 的等级要高,即表示为 $R_1 > R_2$,只有当且仅当:

- 1) $\text{conf}(R_1) > \text{conf}(R_2)$
- 2) $\text{conf}(R_1) = \text{conf}(R_2), \text{sup}(R_1) > \text{sup}(R_2)$
- 3) $\text{conf}(R_1) = \text{conf}(R_2), \text{sup}(R_1) > \text{sup}(R_2)$,但 R_1 左侧的属性值少于 R_2 的左侧。另外,当且仅当 P 是 P' 的一个子集时,一个规则 $R_1: P \rightarrow c$ 被认为是一个关于 $R_2: P' \rightarrow c'$ 的一般规则。

这种方法确实能比较有效地对规则进行分级,但也存在着一些缺陷,例如:当对一个特殊类,最小支持度设为1%或更低的时候,将很可能发生一些有很高置信度参数的规则却被很少的实例确认的情况。这就是为什么以很小的支持度来寻找关联规则将是很危险的,即使这些规则可能看起来很有意义。在本文中,我们使用一种扩展的 χ^2 检验来对规则进行

分级。

传统的 χ^2 检验是一个广泛用于独立性或相关性检验的方法^[11]。实际上,它是基于观测频繁值和期望频繁值之间的比较。对每个规则 $X \Rightarrow Y$ 和训练数据集 D ,规则 $X \Rightarrow Y$ 的 χ^2 值可以用计算公式表达:

$$\chi^2 = \frac{(a_{11}a_{22} - a_{12}a_{21})^2 |D|}{(a_{11}a_{12})(a_{21}a_{22})(a_{11}a_{21})(a_{12}a_{22})}$$

然而,简单地使用传统的 χ^2 值^[13]将对行的分布与列的分布近似的情况有利。因此,我们提出扩展的卡方分布来弥补这个不足:

首先,按照它们的相关性,给出局部最大 χ^2 和全局最大 χ^2 的定义:

给定数据集 D 和类标 Y , $|D|$ 和 Y 的支持数则是固定的。 $g \max(\chi^2)$ 是最大 χ^2 值,意味着规则 $X \Rightarrow Y$ 可能产生,而 $l \max(\chi^2)$ 是在 X 的支持数固定的情况下的最大 χ^2 值。

定义8

$$l \max(\chi^2) = \frac{(n_1 n_2)^2 |D|}{(a_{11}a_{12})(a_{21}a_{22})(a_{11}a_{21})(a_{12}a_{22})}$$

$$n_1 = \min(\min(a_{11} + a_{12}, a_{21} + a_{22}), \min(a_{11} + a_{21}, a_{12} + a_{22}))$$

$$n_2 = \min(\max(a_{11} + a_{12}, a_{21} + a_{22}), \max(a_{11} + a_{21}, a_{12} + a_{22}))$$

即:当给定 X 的支持数,局部最大 χ^2 值在对期望频繁值偏离最大时取到。

定义9 $g \max(\chi^2) = |D|$

证明 为了不失一般性,我们设 $a_{11} + a_{12} \geq a_{21} + a_{22}$; $a_{11} + a_{21} \geq a_{12} + a_{22}$,则有:

$$n_1^2 = (\min(a_{21} + a_{22}, a_{12} + a_{22}))^2 \leq (a_{21} + a_{22})(a_{12} + a_{22})$$

$$n_2^2 = (\min(a_{11} + a_{12}, a_{11} + a_{21}))^2 \leq (a_{11} + a_{12})(a_{11} + a_{21})$$

因此有: $l \max(\chi^2) \leq |D| = g \max(\chi^2)$ 。

当 $a_{21} + a_{22} = a_{12} + a_{22}$ 和 $a_{11} + a_{12} = a_{11} + a_{21}$ 时等式成立,即所有行的分布等于所有列的分布。

现在我们使用CMAR^[4]算法中的例子来说明当简单的选取 χ^2 值作为关联分类的意义度指数时所产生的问题,这也正是我们设计新指数的原因,即扩展的 χ^2 。

例:在一个信用卡申请批准案例中,产生了三个规则:

$$r_1: \text{job} = \text{no} \Rightarrow \text{rejected}(\text{sup} = 30, \text{conf} = 60\%)$$

$$r_2: \text{education} = \text{university} \Rightarrow \text{approved}(\text{sup} = 199, \text{conf} = 99.5\%)$$

$$r_3: \text{number. of. children} > 4 \Rightarrow \text{rejected}(\text{sup} = 2, \text{conf} = 100\%)$$

三个规则的分布如表1所示。三个规则的 χ^2 值分别是88.7, 33.4和18.1,局部最大 χ^2 值分别为287.2, 37.0和18.1。这证明了 χ^2 值对总体行分布与总体列分布近似的情况有利。对一个无工作受过大学教育的客户,如果规则选取是仅基于 χ^2 值的话,参照规则 r_1 他的申请将被拒绝。但是,由于更高的支持度和置信度, r_2 直观地好于 r_1 。并且,尽管 r_3 的支持度非常低,但 r_3 有100%的置信度。 r_3 的意义度基于 χ^2 的话则被低估了。

由于 χ^2 值偏向于行总体分布,为了把它调整得适应面更广,改进了文献[9]提出的扩展的 χ^2 检验的新指数,表示为

$\exp(\chi^2)$ 。

表 1 三个规则的分布描述

r_1	Apply	Reject	Total
Job = yes	438	32	470
Job = no	12	18	30
Total	450	50	500
r_2	Apply	Reject	Total
Ed = uni	199	1	200
Ed \neq uni	251	49	300
Total	450	50	500
r_3	Apply	Reject	Total
Child ≤ 4	450	48	498
Child > 4	0	2	2
Total	450	50	500

定义 10 由上可得:

$$\exp(\chi^2) = \left(\frac{|D|}{l \max(\chi^2)} \right)^a \chi^2$$

$$\frac{\exp(\chi^2)}{\chi^2} = \left(\frac{g \max(\chi^2)}{l \max(\chi^2)} \right)^a = \left(\frac{|D|}{l \max(\chi^2)} \right)^a$$

其中 a 的取值为 0 到 1。参数 a 用来控制全局和局部最大 χ^2 值的影响力,并对应不同的分类问题加以调整。当 $a = 0.5$ 时,这三个规则的扩展的 χ^2 值分别为 117.0, 136.1 和 95.1,这三个值相对之前要合理。可以看到,扩展的 χ^2 值对 $S(Y)$ 或 D 大小的变化是很敏感的。而且,对那些高置信度低支持度的规则,扩展的 χ^2 值在评估它们的优先级时表现得更加谨慎而合理。

在这里 ERAC 算法采用 $\exp(\chi^2)$ 值来对规则进行分级,

$\exp(\chi^2)$ 值越大的规则其优先级越高,当两个规则的 $\exp(\chi^2)$ 值相同时,则采用与 CBA 算法相同的规则分级技术对两者进行分级。

2.3 建立分类器

一个规则只有至少覆盖了一个训练实例的情况下它才是有意义的。在规则被生成且分级以后,建立分类器将参照训练数据集对每个规则依次测试,以此来剔除对一个实例都不能进行分类的规则。在评价程序的每个步骤中,所有已被当前规则正确分类的行将被从训练数据集中删除。任何时候,当一个规则不能对数据的某一行进行分类的话,它将被替换出规则集,因为有一个更高等级的规则已经对这个实例进行了正确的分类。这个进程确保了只有高置信度的规则才能留在 ERAC 的分类器中。

3 实验结果

实验使用的主机配置为:操作系统为 Windows 2000, CPU 为 P4 2.4 GHz, 内存为 512 MB。我们一共选取了 C4.5、PART、BN (Bayesnet)^[10]、CBA 这 4 个算法与 ERAC 算法来进行实验比较,其中 C4.5、PART、BN 这 3 个算法使用 WAKE^[12] 系统中的算法进行测试,CBA 算法使用 VC++ 编写的软件实现,ERAC 算法使用 Java 在 NetBeans 5.5 中编程实现。所有结果都是由交叉验证法验证得出,CBA 和 ERAC 算法关于支持度和置信度阈值使用关联分类算法实验时的常用标准,也就是 $Min_{sup} = 1\%$, $Min_{conf} = 50\%$ 。

实验使用的是 UCI 数据库中的 20 个标准数据集,数据集的相关信息和最终的结果见表 2。

表 2 实验数据相关描述和分类结果

样本名称	样本数目	属性数目	类别数目	C4.5/%	PART/%	BN/%	CBA/%	ERAC/%
Anneal	898	38	6	92.09	94.88	92.54	97.13	97.78
Austral	690	14	2	85.51	85.51	85.51	86.53	87.41
Breast cancer	699	11	2	94.42	94.28	97.28	95.13	94.27
Crx	690	15	2	85.80	85.94	86.52	85.24	87.10
Heart disease	270	14	2	80.01	83.70	83.70	81.87	83.70
Hepatic	155	19	2	81.94	79.35	83.23	81.82	81.37
Horse	368	22	2	85.33	85.01	80.43	81.53	85.88
Hypo	958	10	2	99.21	97.35	96.47	98.87	99.10
Iono	846	19	4	89.17	91.45	88.89	92.30	91.97
Labor	699	11	2	78.95	80.70	91.23	86.33	88.33
Led7	3 200	7	10	73.34	73.56	73.16	69.54	70.70
Lymph	148	18	4	77.03	79.05	87.84	79.31	81.65
Pima	768	8	2	71.09	70.70	74.74	73.67	73.41
Tic-tac	958	9	2	83.72	92.59	69.94	98.86	100.00
Thyroid disease	3 772	12	6	99.71	99.60	98.75	98.69	98.90
Vehicle	846	18	4	71.04	71.87	61.35	68.03	68.10
Vote	435	17	2	77.34	81.27	74.92	99.77	99.77
Waveform	5 000	21	3	77.04	78.00	80.82	81.28	82.32
Wine	178	13	3	94.38	91.01	98.31	94.96	94.96
Zoo	101	16	7	92.21	93.07	93.07	97.78	97.10
平均精度				84.47	85.44	84.94	87.43	88.19

从表 2 中可以看出,基于关联规则的分类算法 CBA 和 ERAC 的平均准确率均超过了 C4.5 等其他 3 个算法,而 ERAC 的 88.19% 略优于 CBA 算法的 87.43%。

我们从 20 个数据集中取出 CBA 算法与 ERAC 算法性能基本相当的十个数据集,计算了两个算法在交叉验证中分类器包含关联规则数目的平均值。在分类器性能基本相当的情况下,分类器包含规则数目越小,算法最后执行分类预测

的效率也要越高。

从图 1 中我们可以看出,ERAC 算法在分类准确度与 CBA 算法相同的情况下,分类器所含的规则数目明显少于 CBA 算法,差距最大的时候 (Austral), CBA 分类器所含规则数目是 ERAC 分类器规则数目的 5 倍。

为了说明分类器规则数目对算法运行效率的影响,我们又取出 5 个数据集测试两个算法的运行时间进行比较,结果

如图2。从图中可以看出,在运行时间这一指标上,ERAC 算法的训练时间明显小于 CBA 算法,一般只有 CBA 算法所耗时间的一半左右。

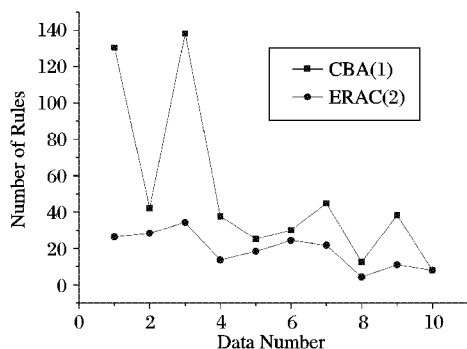


图1 分类器所含规则数目

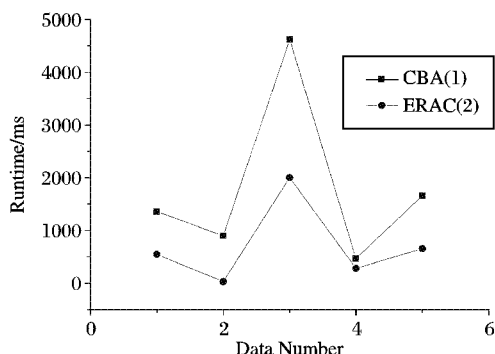


图2 算法运行时间

4 结语

本文提出了一种新的用于多类别分类的方法:ERAC 关联分类算法。ERAC 与传统的关联分类法又很多不同之处,主要在于:1)使用了新的算法来发掘规则,它只需要对训练数据扫描一次;2)介绍了一种剪除多余规则和确保只保留高效规则用于分类的有效分级技术;3)把频繁项集的发现和规则产生结合为一步,节省了存储空间和运行时间的开支。对基于取自 UCI 机器学习数据集中的 20 个数据集的运行显示了我们的方法是有效和一致的,并比 C4.5、PART、BN 和 CBA 算法有着更高的分类准确率。

(上接第 1680 页)

有界的线性空间基矢量。对于空间基,取多项式的基底,一般地讲不能满足回归稳健性,只是部分地保证回归稳健。但是对于本文的光滑问题,在将时间正规化单位化后, x 有界, $|x| \leq k$, 基矢量也有界。这样采用二次多项式线性回归空间的稳健光滑则保证了回归的稳健性。

4 结语

经典的最小二乘回归数据光滑方法计算起来十分简单,为矫正某数值,只要将 $2k+1$ 个数据各乘以固定数值,再取和即可。这种回归假设 $2k+1$ 个点的残差服从 iid 分布,在此基础上进行极大似然估计,获得最佳回归结果。但是误差的理想化分布和实际情况是有差异的,由于客观或人为原因,使得数据存在异常点。由于数据的偏离,使得最佳回归方法的效果产生了差异,这种回归是不稳健的。为使数据光滑具有稳健性,本文采用 M 稳健二次多项式回归,给出了该回归的影响函数 IF,指出了影响 IF 的几个因素,并从 IF 分析论证了该方法的全稳健性。

参考文献:

- [1] AGRAWAL R, IMIELINSKI T, SWAMI A. Mining association rules between sets of items in large databases [C]// Proceedings of the ACM SIGMOD Conference on Management of data. Washington, DC: ACM Press, 1993: 207–216.
- [2] BOJARCZUK C C, LOPES H S, FREITAS A A. Genetic programming for knowledge discovery in chest-pain diagnosis [J]. IEEE Engineering in Medicine and Biology Magazine, 2000, 4(19): 38–44.
- [3] LIU B, HSU W, MA Y. Integrating classification and association rule mining [C]// Proceedings of the KDD. New York: [s. n.], 1998: 80–86.
- [4] LI W, HAN J. CMAR: Accurate and efficient classification based on multiple class-association rules [C]// Proceedings of the 2001 International Conference on Data Mining. [S. l.]: IEEE Press, 2001: 369–376.
- [5] BARALIS, GARZA P. A lazy approach to pruning classification rules [C]// Proceedings of the IEEE 2002 International Conference on Data Mining. Japan: IEEE Press, 2002: 35–42.
- [6] YIN X, HAN J. CPAR: Classification based on predictive association rules [C]// Proceedings of the Third SIAM International Conference on Data Mining. San Francisco: [s. n.], 2003.
- [7] JANSSENS D, LAN Y, WETS G. Empirically validating an adapted classification based on associations algorithm on UCI DATA [C]// Applied Computational Intelligence Proceedings of the 6th International FLINS Conference. Blankenberge, Belgium: World Scientific Publishing, 2004: 167–172.
- [8] SUN Y, WANG Y, ANDREW. Boosting an Associative Classifier [J]. IEEE Transactions on Knowledge and Data Engineering, 2006, 18(7).
- [9] YU L, CHEN G, JANSSENS D, WETS G. Dilated chi-square: A novel interestingness measure to build accurate and compact decision list [C]// Proceedings of the international conference on intelligent information processing. Beijing: [s. n.], 2001: 233–237.
- [10] TAN P N, TAN M. S, KUMAR V. Introduction to data mining [M]. Redwood: Addison Wesley, 2005.
- [11] MILLS F. Statistical methods [M]. London: Pitman M, 1955.
- [12] WITTEN I H, FRANK E. Data mining: Practical machine learning tools and techniques with Java implementations [M]. 2nd ed. San Francisco, CA: Morgan Kaufmann, 2005.
- [13] 印鉴, 谭焕云. 基于 χ^2 统计量的 KNN 文本分类算法 [J]. 小型微型计算机系统, 2007, 6(6): 1094–1097.

参考文献:

- [1] 何红丽, 任朴舟, 田伟峰, 等. 修正的 MT 滤波清除算法在飞行试验数据处理中的应用研究 [J]. 测控技术, 2006, 25(4): 11–13.
- [2] 李根强. 基于三次样条插值的采样数据光滑曲线形成法 [J]. 计算技术与自动化, 2001, 20(1): 59–62.
- [3] 徐国良, 张琴, 刘丹. 带噪声散乱数据的光滑曲面重构—变分水平集方法 [J]. 计算机辅助设计与图形学学报, 2007(7): 840–848.
- [4] 罗双华, 玄海燕. 缺失数据下半参数回归模型的局部线性光滑 [J]. 兰州理工大学学报, 2007, 33(5): 151–155.
- [5] 路威, 余旭初, 刘娟. 高光谱遥感数据三次光滑样条滤波 [J]. 测绘学院学报, 2005, 22(1): 11–13, 16.
- [6] HUBER P J. Robust statistics [M]. New York: John Wiley Press, 1988.
- [7] 张世英. 测量实践的数据处理 [M]. 北京: 科学出版社, 2000.
- [8] 闫章更, 魏振军. 试验数据的统计分析 [M]. 北京: 国防工业出版社, 2001.
- [9] 刘利生. 外测数据事后处理 [M]. 北京: 国防工业出版社, 2000.