文章编号:1001 - 9081(2007)04 - 0892 - 03

关联规则相关性的度量

郭俊芳[1],谢益武[1],周生宝[2]
(1. 大连海事大学 计算机科学与技术学院,辽宁 大连 116026; 2. 大连海事大学 数理系,辽宁 大连 116026)
(zhouguo@ newmail. dlmu. edu. cn)

摘 要:用 Apriori 算法生成的关联规则包含有无用规则,甚至误导规则。为了使生成的规则更有效,引入了统计学中的卡方检验从统计意义上检验规则是否关联,并找到卡方检验值与相关系数的数量关系,实现了两种方法的统一,并用基于相关系数的算法去生成关联规则。

关键词:关联规则;相关度;卡方检验;相关系数
中图分类号: TP311.13;TP301.6 **文献标识码**:A

Measurement of association rules correlation

GUO Jun-fang[1], XIE Yi-wu[1], ZHOU Sheng-bao[2]
(1. *College of Computer Science and Technology, Dalian Maritime University, Dalian Liaoning* 116026, *China*;
2. *Department of Mathematics and Physics, Dalian Maritime University, Dalian Liaoning* 116026, *China*)

Abstract: Association rules generated by Apriori algorithm includes some useless and even misleading rules. To gain more effective rules, a statistical criteria, Chi-Squared was used to measure the associations, furthermore the quantitative relations between the Chi-Squred test and correlation coefficient was found out, and two measures was united to generate association rules with correlation coefficient.

Key words: association rules; correlation; chi-squared test; correlation coefficient

0 引言

在支持度—置信度框架下,关联规则是数据项同时满足最小支持度阈值(minsup)和最小置信度阈值(minconf)的规则。但用此框架产生的规则有些是多余的,有些甚至具有很强的误导性。

某超市一段时间内形成 1 000 条交易记录,同时购买牛奶、可乐,只买牛奶,只买可乐,两者都不买的数据记入表 1。此表在统计学中称为列联表。

考察买牛奶与买可乐的关系:设 minsup = 0.3, min$conf$ = 0.6。

表 1 牛奶与可乐列联表

	可乐	$\overline{\text{可乐}}$	行总计
牛奶	400	200	600
$\overline{\text{牛奶}}$	350	50	400
列总计	750	250	1 000

sup(牛奶 ⇒ 可乐) = 0.4, $conf$(牛奶 ⇒ 可乐) ≈ 0.66 > min$conf$

根据支持度—置信度框架得出 是强关联规则,即买牛奶的人有 66% 的可能性会买可乐,增加牛奶的销量就会刺激可乐的销量。但事实果真如此吗?

再看不买牛奶与买可乐的关系:

sup($\overline{\text{牛奶}}$⇒ 可乐) = 0.35, $conf$($\overline{\text{牛奶}}$⇒ 可乐) = 0.875 > min$conf$

因此$\overline{\text{牛奶}}$⇒ 可乐也是一条强关联规则,即减少牛奶的销量会增加可乐的销量,显然这与规则牛奶 ⇒ 可乐自相矛盾。

产生上述现象是因为规则的置信度有一定的欺骗性,$conf(A \Rightarrow B)$ 只是在给定 A 的情况下 B 出现的条件概率,并没有考虑 B 在整个数据集中出现的随机概率 $sup(B)$。如果 B 的随机概率本身就很大,那么置信度大的规则就不一定是强蕴涵关系。所以单凭置信度不能将强蕴涵的关联关系从随机关系中分离出来。上例可乐的随机概率 sup(可乐) = 0.75 > $conf$(牛奶 ⇒ 可乐) ≈ 0.66,这意味着在买牛奶的前提下买可乐的概率比随机情况还少9%,显然是错误的。而 $conf$($\overline{\text{牛奶}}$⇒ 可乐) = 0.875 > sup(可乐) = 0.75 是正确的强蕴涵规则,因为不买牛奶会买可乐的概率从随机情况下提高 12.5%。这时可以说牛奶和可乐是负关联的,一个出现会减少另一个出现的概率。另一种情况 $sup(A) = 1$, $sup(B) = 0.7$, $conf(A \Rightarrow B) = 0.7$, $conf(B \Rightarrow A) = 1$,但显然 A 与 B 没有关系是独立的,也即置信度为 1 的规则并不一定是强关联规则。因此需要引入其他方法去度量两项间是否有关联及关联程度以减少弱关联规则,负关联规则。严格说关联与不独立的概念是一致的,不同于相关。关联包括各种关系而相关只指线性关系。本文介绍了卡方检验(独立性检验),发现了对于二态变量关联性与相关性的关系,得出可通过度量相关性去获得关联规则。

1 卡方检验

针对项集间的关联关系许多学者、专家进行了研究。Piatetsky-Shapiro 提出了 P-S 兴趣度, $interest = sup(A \Rightarrow B) - sup(A)sup(B)$。在文献[5]中作者对其进行了改进,综合考虑了用户主观偏好、规则准确度、规则相关度及兴趣度。文献[7]中作者提出用有效度代替置信度,$validity = p(AB) - p(\bar{A}B)$。文献[4]中作者提出匹配度,$match = conf(A \Rightarrow B) - conf(\bar{A} \Rightarrow B)$。以上方法均能在一定程度上减少无用规则产生,但不能准确度量关联程度,因而产生的规则仍有一定误导性。

收稿日期:2006 - 10 - 09;**修订日期**:2006 - 12 - 24

作者简介:郭俊芳(1980 -),女,山西大同人,硕士研究生,主要研究方向:数据库与信息系统; 谢益武(1965 -),男,安徽桐城人,教授,主要研究方向:数据库与信息系统; 周生宝(1979 -),男,山西闻喜人,硕士研究生,主要研究方向:组合数学.

文献[1]通过比较 $conf(A\Rightarrow B)$ 与 $sup(B)$ 来度量 A 与 B 的相关性:

$$corr_{A,B}=\begin{cases}(1,+\infty), & \text{条件概率 } p(B|A) \text{ 大于 } B \text{ 的随机概率 } p(B),A,B \text{ 正相关}\\ 1, & p(AB)=p(A)p(B),A \text{ 与 } B \text{ 相互独立}\\ [0,1), & \text{条件概率 } p(B|A) \text{ 小于 } B \text{ 的随机概率 } p(B),A,B \text{ 负相关}\end{cases}$$

上例中 $corr_{\text{牛奶,可乐}}=\dfrac{400}{600\times 750/1\,000}=0.88$,所以牛奶与可乐是负相关的。

然而,有的事件没有如此清晰明确。比如若 $p(AB)$ 与 $p(A)p(B)$ 很接近,那么 A 与 B 是独立呢还是关联呢?以什么标准衡量呢?且 $corr_{AB}$ 值也不能准确反映相关程度,比如有两组数据,第一组 $p(AB)=0.6,p(A)=0.6,p(B)=0.6$;第二组 $p(AB)=0.8,p(A)=0.8,p(B)=0.8$,它们的 $corr$ 值不相等,而事实上两组数据中 A 与 B 都完全相关。用 P-S 兴趣度度量也有类似的问题。所以需要从统计意义上用卡方检验确定两项是否关联。

为了阐述方便先做出 A,B 列联表。n_{ij} 表示状态组合 (i,j) 发生的频数。

卡方检验的思想是:若 A 与 B 独立,则 $p(AB)=p(A)p(B)$,即 $\dfrac{n_{AB}}{n}=\dfrac{n_{A.}}{n}\cdot\dfrac{n_{.B}}{n}$,即 $n_{AB}=\dfrac{n_{A.}n_{.B}}{n}$(其余三种组合也同理)。若 A 与 B 关联,则 n_{AB} 与 $\dfrac{n_{A.}\cdot n_{.B}}{n}$ 就有差距。通过:

表 2　A,B 列联表

	B	$\bar{B}$	$\sum_{row}$
A	n_{AB}	$n_{A\bar{B}}$	n_A
$\bar{A}$	$n_{\bar{A}B}$	$n_{\bar{A}\bar{B}}$	$n_{\bar{A}}$
$\sum_{col}$	$n_{.B}$	$n_{.\bar{B}}$	n

$$Q^2=\frac{\left(n_{AB}-\frac{n_{A.}\cdot n_{.B}}{n}\right)^2}{\frac{n_{A.}\cdot n_{.B}}{n}}+\frac{\left(n_{A\bar{B}}-\frac{n_{A.}\cdot n_{.\bar{B}}}{n}\right)^2}{\frac{n_{A.}\cdot n_{.\bar{B}}}{n}}+\frac{\left(n_{\bar{A}B}-\frac{n_{\bar{A}.}\cdot n_{.B}}{n}\right)^2}{\frac{n_{\bar{A}.}\cdot n_{.B}}{n}}+\frac{\left(n_{\bar{A}\bar{B}}-\frac{n_{\bar{A}.}\cdot n_{.\bar{B}}}{n}\right)^2}{\frac{n_{\bar{A}.}\cdot n_{.\bar{B}}}{n}}\tag{1}$$

来计算它们间的差距。此统计量称为 Pearson χ^2 统计量。经证明当 A 与 B 独立时,Q^2 渐进服从 $\chi^2(1)$ 分布。科克伦曾提出一个标准,在 n 大于 20,各组理论频数小于 5 的不超过 20% 时(小于5的格要与相邻格进行合并),Q^2 服从 $\chi^2(1)$ 分布。这样可根据统计学中的假设检验理论来考察两项是否关联。给定一个显著水平 α(常为0.01,0.05,0,1等),首先假设 A 与 B 独立,算出 Q^2,然后,查 $\chi^2(1)$ 分布表得分位数 $\chi^2_{1-\alpha}(1)$。若 $Q^2>\chi^2_{1-\alpha}(1)$ 这一小概率事件发生了,则在 $1-\alpha$ 的级别上不接受独立假设,换句话说 A 与 B 关联的置信度为 $1-\alpha$。否则若 $Q^2<\chi^2_{1-\alpha}(1)$,则假设为真,A 与 B 独立。

上例中令 $\alpha=0.05$,则:

$$Q^2=\frac{(400-450)^2}{450}+\frac{(200-150)^2}{150}+\frac{(350-300)^2}{300}+\frac{(50-100)^2}{100}$$
$$=55.56>\chi^2_{0.95}(1)=3.841$$

所以从统计意义上说牛奶与可乐不独立,置信度为 95%。

$$corr_{A,B}=\frac{p(B|A)}{p(B)}=\frac{p(AB)}{p(A)\cdot p(B)}$$

对 2×2 列联表文献[9]推导出统计量:

$$Q^2=\frac{n\times(n_{AB}\cdot n_{\bar{A}\bar{B}}-n_{A\bar{B}}\cdot n_{\bar{A}B})^2}{n_{A.}n_{\bar{A}.}n_{.B}n_{.\bar{B}}}\tag{2}$$

文献[8]指出二态变量 A 与 B 的相关系数为:

$$\rho_{AB}=\frac{n_{AB}\cdot n_{\bar{A}\bar{B}}-n_{A\bar{B}}\cdot n_{\bar{A}B}}{\sqrt{n_{A.}n_{\bar{A}.}n_{.B}n_{.\bar{B}}}}\tag{3}$$

$$\rho_{AB}=\begin{cases}(0,1] & A \text{ 与 } B \text{ 正相关}\\ 0 & A \text{ 与 } B \text{ 独立}\\ [-1,0) & A \text{ 与 } B \text{ 负相关}\end{cases}$$

且 $|\rho_{AB}|$ 值越大,相关程度越强。

比较(2),(3)式有 $\rho=\sqrt{\frac{1}{n}Q^2}$,即说明了对于固定的 n,两个二态变量的不独立与相关性大的概念是一致的。所以我们完全可通过计算 A 与 B 的相关系数来确定两项是否关联及关联程度。这时关联与相关的意义是一致的。上例计算得:

$$corr_{\text{牛奶可乐}}=\frac{400\times 50-200\times 350}{\sqrt{600\times 400\times 750\times 250}}=-\frac{1}{3\sqrt{2}}\approx -0.236$$

故牛奶与可乐负相关。对于形如 $AB\Rightarrow C$ 的三项或更多项规则,同样可以先计算出 $ABC,AB\bar{C},A\bar{B}C,\bar{A}BC$ 的理论频数后由(3)式去判定是否相关。

2　算法思想

1) 给定 minsup,min$corr$,(min$corr\geqslant\sqrt{\frac{\chi^2(1-\alpha)(1)}{n}}$ 是依据规则成立的置信度水平和主观要求而设定的),由 Apriori 算法或其改进算法先生成频繁项集集合 $L=L_k$;

2) 只需利用第一部分产生的候选项支持度计数 n_A,n_B,n_{AB} 去计算 ρ_{AB},而不用计算其他的理论频数。推导如下:设项 i 的支持度计数用 C_i 表示,则:

$$n_{A.}=C_A,n_{\bar{A}.}=n-C_A,n_{.B}=C_B,n_{.\bar{B}}=n-C_B,n_{AB}=C_{AB},$$
$$n_{A\bar{B}}=C_A-C_{AB},n_{\bar{A}B}=C_B-C_{AB},n_{\bar{A}\bar{B}}=n-C_A-C_B+C_{AB}$$

$$\rho_{AB}=\frac{n_{AB}\cdot n_{\bar{A}\bar{B}}-n_{A\bar{B}}\cdot n_{\bar{A}B}}{\sqrt{n_{A.}n_{\bar{A}.}n_{.B}n_{.\bar{B}}}}=\frac{C_{AB}\cdot(n-C_A-C_B+C_{AB})-(C_A-C_B)\cdot(C_B-C_{AB})}{\sqrt{C_A\cdot(n-C_A)\cdot C_B(n-C_B)}}=\frac{nC_{AB}-C_AC_B}{\sqrt{C_A\cdot(n-C_A)\cdot C_B(n-C_B)}}\tag{4}$$

若 $C_A=n$ 或 $C_B=n$,则显然 A 与 B 是独立的,就不再需要计算 ρ_{AB}.

观察(4)式的分子与 P-S 兴趣度定义可发现:分子 $=n^2\cdot interest$,但分子除以分母部分就修正了用 $interest$ 度量前述两组数据时兴趣度不一致的问题,从而得出他们具有相同相关度的正确结论。

3) 比较 ρ_{AB} 与 min$corr$,若 $\rho_{AB}\geqslant$ min$corr$ 称 $A\Rightarrow B$ 或 $B\Rightarrow A$ 为强关联规则,否则舍弃。算法并不需要构造列联表或其他复杂计算,所以在时间和空间上都非常高效,且易于编程实现。

(下转第 896 页)

取某些值的情况下,K-NFP 的分类性能甚至比 NFP 还好,如 $K=3$、4、5 等,而在这些情况下,由于建立特征平面的数量大大减少,显然其实时性能也要明显优于 NFP。另一方面,不论是 NFL 与 NFP,还是 K-NFL 与 K-NFP,其分类性能较之于 NN 都具有较大的优势。

3 结语

研究了 NFL 与 NFP 分类器的分类机理,并用一种新的搜索策略对它们进行了修正,得出了 K-NFL 与 K-NFP 分类器。用三种不同类型飞机的实测回波数据进行了分类实验,结果表明基于主分量分析特征提取,K-NFL 与 K-NFP 在实时性能上要明显优于原来的 NFL 与 NFP,当 K 取一定数值时,识别性能仍保持较高水平(对 K-NFP 而言甚至还要优于 NFP)。下一步的工作将根据不同的特征提取方法与不同分类器组合,其识别性能不同的观点,从特征提取后数据分布的角度出发,对特征提取与分类器的结合做进一步研究。

参考文献:

[1] COVER TM, HART PE. Nearest neighbor pattern classification[J]. IEEE Transactions on Information Theory, 1967, 13: 21 - 27.

[2] BARTLETT MS, MOVELLAN JP, SEJNOWSKI TJ. Face recognition by independent component analysis[J]. IEEE Transactions on Neural Networks, 2002, 13(6): 1450 - 1464.

[3] DOMINIQUE V, HERVE A, ALICE JO, *et al*. Connectionist models of face processing: A survey[J]. Pattern Recognition, 1994, 27(9): 1209 - 1230.

[4] FRIEDMAN N, GEIGER D, GOLDSZMIDT M. Bayesian network classifier[J]. Machine Learning, 1997, 29: 131 - 163.

[5] VAPNIK N. The nature of statistical learning theory[M]. New York, Springer-Verlag, 1995. 1 - 188.

[6] LI SZ, LU JW. Face recognition using the nearest feature line method[J]. IEEE Transactions on Neural Networks, 1999, 10(2): 439 - 443.

[7] CHIEN JT, WU CC. Discriminant waveletfaces and nearest feature classifiers for face recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002, 24(12): 1644 - 1649.

[8] ZHENG WM, ZHAO L, ZOU CR. Locally nearest neighbor classifiers for pattern classification[J]. Pattern Recognition, 2004, 37(6): 1307 - 1309.

[9] NOVAK LM, OWIRKA GJ. Radar target recognition using an eigenimage approach[A]. IEEE International Radar Conference[C]. 1994. 129 - 131.

(上接第 893 页)

3 算法实验验证

为验证算法有效性,用 Java 语言编程对某超市事务数据库 1000 条数据进行了测试。结果发现算法能有效去除负关联和和弱关联规则。且在给定显著水平 α 时,置信度越小,两个算法生成的规则数量差别越大,随着置信度的提高差别缩小。最小置信度给定时,显著水平越小生成规则数量差别也小,随着 α 增大,差别变大。

以下给出试验中的 60 个事务,每个事务有五项,每一项已经进行了 0 - 1 标准化。并对比给出两个算法生成的规则。

表 1 60 个事务的数据

00110	01011	01011	11011	01111	10001
01111	11111	10001	01000	00110	01011
01011	10110	01111	10000	01111	11111
10001	10010	01001	00100	00110	10010
01011	01110	01011	11001	11011	10001
01111	00101	10000	01011	01111	10101
11111	00000	10001	10010	01001	00101
10010	01110	11001	10001	00101	01011
10101	00000	10010	00101	10010	01110
11001	10001	00101	01011	10100	00001

用 Apriori 算法生成的实验数据如下:

设 $\min sup = 40\%$, $\min conf = 60\%$,则规则如下:

$A \Rightarrow E; B \Rightarrow D; D \Rightarrow B; B \Rightarrow E; E \Rightarrow B; C \Rightarrow D; C \Rightarrow E; D \Rightarrow E; BD \Rightarrow E; B \Rightarrow DE; DE \Rightarrow B; D \Rightarrow BE; BE \Rightarrow D;$

用基于相关系数算法得到的实验数据如下:

设 $\alpha = 0.05$,则 $\min corr \geqslant \sqrt{\frac{x^2_{(1-\alpha)}(1)}{n}} \approx 0.2530$

若设 $\min sup = 40\%$, $\min corr = 0.2530$ 则数据如下:$\rho_{BD} = 0.4726$; $\rho_{BE} = 0.4009$; $\rho_{E\&BD} = 0.3394$; $\rho_{B\&DE} = 0.7311$; $\rho_{D\&BE} = 0.4247$; $\rho_{CD} = 0.1529$; $\rho_{AE} = -0.0711$; $\rho_{CE} = -0.0478$; $\rho_{DE} = -0.1421$。

算法检测出在 Apriori 算法生成的 13 个规则中有 3 个负关联规则,分别为 $A \Rightarrow E; C \Rightarrow E; D \Rightarrow E$;一个弱关联规则 $C \Rightarrow D$。所以基于相关系数得到的关联规则能有效避免弱关联和负关联规则。

4 结语

本文通过实例分析了支持度—置信度框架的不足,主要在于置信度并不能很好地反应项集间的关系,从而产生冗余,甚至是错误的规则。为了准确清晰地度量关联程度,引入了统计学中的卡方检验,并找到了它和相关系数的关系。经证明 $Q^2 = n(\rho_{AB})^2$,这样可通过计算相关系数去排除弱关联及负关联规则,同时本文也给出了可实现的算法。

参考文献:

[1] HAN JW, KAMBER M. 数据挖掘概念与技术[M]. 范明, 等译. 北京: 北京机械工业出版社, 2001.

[2] 夏火松. 数据仓库与数据挖掘技术[M]. 北京: 科学出版社, 2004. 157 - 160.

[3] 徐勇, 周森鑫. 一种改进的关联规则挖掘方法研究[J]. 计算机技术与发展, 2006, 16(3): 77 - 79.

[4] 伊卫国, 卫金茂, 王名扬. 关联规则挖掘方法的改进[J]. 东北师大学报, 2006, 38(2): 15 - 19.

[5] 杨建林, 邓三鸿, 苏新宁. 关联规则兴趣度的度量[J]. 情报学报, 2003, 22(4): 419 - 424.

[6] XU Y, ZHOU SX, GONG JH. Mining association rules with new measure criteria[A]. Proceedings of the Fourth International Conference on Machine Learning and Cybernetics[C]. Guangzhou, China, 2005.

[7] 罗可, 吴杰. 怎样获得有效的关联规则[J]. 小型微型计算机系统, 2002, 23(6): 1711 - 1733.

[8] 孙文爽, 陈兰祥. 多元统计分析[M]. 北京: 高等教育出版社, 1994. 341 - 343.

[9] 梅长林, 周家良. 实用统计方法[M]. 北京: 科学出版社, 2002. 198 - 200.

[10] 耿修林, 谢兆茹. 应用统计学[M]. 北京: 科学出版社, 2002. 215 - 218.

[11] CRAWSHAW J, CHAMBERS J. A concise course in A-Level statistics with worked examples[M]. Nelson Thornes Ltd, 2001. 548 - 558.