一种基于相似性的规则集一致性度量的新方法

彭 俊,谢荣传,王大刚,耿 波
(安徽大学 计算机学院,安徽 合肥 230039)

摘 要:规则学习算法通过学习样本产生规则集,如何判断规则集的好坏? 目前规则集的评估标准有很多,如一致性、可测量性和易理解性评估,但它们有各自的缺点。提出一种新的评估规则集方法:相似性度量。这种度量方法可以计算出两个规则集之间的正相似性与负相似性。实验说明这种新的度量方法可以被用来评估规则集间的一致性,并且可以决定使用哪种算法解决某类问题或选择组合分类模型中的基模型。

关键词:正相似性; 负相似性;规则集;一致性
中图分类号:TP311 **文献标识码**:A **文章编号**:1673-629X(2008)11-0073-03

A New Measuring Rule Set Consistency Method Based on Similarity

PENG Jun, XIE Rong-chuan, WANG Da-gang, GENG Bo
(Department of Computer Science and Engineering, Anhui University, Hefei 230039, China)

Abstract: The rule extraction algorithm produces the rule set by learning examples. How to evaluate the rule set? There are several evaluation criteria for rule set at the present time, such as consistency, measurability and comprehensibility, but they have various disadvantages. Proposes a new evaluation measurement: similarity. The evaluation measurement can measure positive similarity and negative similarity between two rule sets. The experiment shows this new measurement can be used to measure consistency between rule sets and decide to choose algorithm or select the elemental model of combination classification model.

Key words: positive similarity; negative similarity; rule set; consistency

0 引 言

规则学习算法可以通过样本训练得到一组分类规则,这组分类规则可以对未知样本进行分类或对未来可能出现的情况做出预测,所以规则学习算法所生成的规则集的分类正确率或预测准确率是选择该算法的关键。由多个规则学习算法在同种训练环境中所生成的多个规则集,必须对其进行评估,判断哪种规则学习算法更适合解决某类已知问题。现在有很多规则集评估标准,如一致性评估、可测量性评估以及易理解性评估[1],这些评估标准从各方面对算法生成的规则集进行评测,例如由神经网络算法生成的分类器的易理解性差[2],而决策树算法生成规则集可以用树形表示,易理解性评估得分高。文中讨论的是一致性相关方面的评估。

收稿日期:2008-02-14
基金项目:安徽省自然科学基金项目(070412051);安徽高校省级重点自然科学研究项目(KJ2007A43)
作者简介:彭 俊(1981-),男,安徽合肥人,硕士研究生,研究方向为数据仓库与数据挖掘;谢荣传,教授,硕士生导师,研究方向为网络与数据库。

笔者提出了一种评估规则集的定义:规则提取算法在同种训练环境中训练产生的多个规则集的相似性,把规则集间的相似性分为正相似性与负相似性,正相似性描述一个规则提取算法在同种训练环境中训练 K 次,产生的 K 个规则集是否具有相同的规则,以及它们对样本正确分类所使用的规则是否相同;负相似性描述了两个规则集之间共同错误分类的相似程度。文中讨论的相似性评估方法可以测量规则集间的相似程度,并给出计算规则集间相似性的算法。下面对该算法进行分析,阐述算法的步骤。此外,还做了一个实验:使用两种规则学习算法,每种规则学习算法分别在多个训练集上训练,得到多个规则集,分别计算每种规则学习算法所产生的多个规则集之间的平均正相似性和平均负相似性,通过计算正相似性确定哪种规则学习算法更适合解决该类问题。并建立两个组合分类模型,测试两个组合分类模型的分类精确度来证明负相似性可以有效地为组合分类模型选择基模型。

1 相似性

相似性评估可以定量地测量两个规则集之间的相

似程度，是一种新的一致性评估方法，传统的一致性评估只能判断算法产生的两个规则集是否完全一致，如两个规则集有相同的规则或者规则集的分类精度一致，那么认为这些算法具有一致性[3]。相似性评估就是对一致性评估的改进，它可以对两个规则集的一致性进行量化表示。

下面对 Johan Huysmans 等提出的测量规则集一致性算法[4]加以改进，提出一个计算规则集间相似性的算法，此算法可以计算不同规则集之间的正相似性与负相似性。算法的主要思想是：如果两个规则集对大多数样本正确分类，并且它们在正确分类时使用相似的规则，那么这两个规则集的正相似性高，而两个规则集对某样本同时错误分类，且错误分类结果一致，如果共同错误分类且误分结果一致的样本数占样本集中被误分样本总数的比重很大，说明两个规则集负相似程度高。

正相似性主要检查两个规则集做出正确分类所使用的规则一致程度，它从正面证明了一个规则提取算法提取的规则集是否适合解决此类问题。

负相似性主要判断两个规则集做出的共同错误分类的一致性，负相似性证明了一个规则学习算法所提取的规则集是否适合该数据集。负相似性的用途很多，例如关于组合分类方法中基分类模型的选择，可以通过计算基分类模型间的负相似性，然后选择负相似程度低的基分类模型组成组合分类器，这样组合分类器的精度更高，更稳定。

用下面的算法去计算规则集 X，Y 的相似性，给出 N 个样本。假设每个规则集中包含的规则是互斥的，也就是说每个样本仅触发一条规则。

(1) 为 X，Y 中的每个规则做初始化设定：分配一个唯一的标记给每个规则。

(2) 对于每个样本，分别利用 X，Y 规则集做出的预测，得到预测结果。

(3) 标记每个样本所触发的规则，并记录每个规则所被触发的次数。

(4) 对每个规则，在其它规则集中，找出与此规则共同分类样本一致且正确次数最多的规则，将被这两条规则一致正确分类的样本的数量记为 N_r^S。

(5) 对每个规则集统计它们共同误分且误分结果一致的样本数 N^F。

(6) 计算 X，Y 中每个规则的 N_r^S 之和，然后除以每个规则正确分类样本数之和，得到规则集 A，B 之间的正相似性。

(7) 共同误分且误分结果一致样本数除 N^F 以样本集中被误分样本总数，得到两个规则集之间的负相似性。

下面通过文中提出的算法计算两个规则集之间的正相似性与负相似性。如图 1 所示：算法生成两个规则集，每个规则集都是将三角形、实心点与空心点分离。为每个规则分配一个唯一符号，规则集 X：$X1\cdots X5$，规则集 Y：$Y1\cdots Y4$。

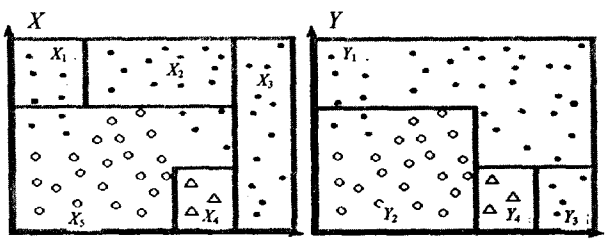

图 1 规则集 X 和规则集 Y

通过图 1 可以看出规则集 X 与规则集 Y 都有误分样本的情况发生，其中规则 $X5$ 将 7 个实心点误分成空心点，规则 $Y2$ 将 3 个实心点误分成空心点，但是它们所误分的样本是不同的，这就说明了两个规则集之间的不一致性，现在来计算规则集 X，Y 两者之间的相似性。

首先计算规则集 X，Y 之间的正相似性：

规则集 X 与规则集 Y 分别对每个样本进行分类。找出每个规则所正确分类的样本数，例如规则 $X1$ 对 5 个样本进行正确分类。然后寻找规则集 Y 与此规则共同分类样本一致正确次数最多的那个规则。因为被 $X1$ 分类的 5 个实心点全部被 $Y2$ 分类，且分类结果一致正确，所以 $N_{X1}^S = 5$，$Y4$ 对 3 个样本分类，发现这 3 个三角形同时被 $X4$ 分类，并且分类结果一致正确，所以 $N_{y4}^S = 3$。按照这种方法，计算 X，Y 规则集中的每个规则的 N_r^S，将其写入表 1。

表 1 计算规则集 X，Y 之间的正相似性

规则	规则正确分类样本数	与其它规则共同分类且结果一致正确样本数	共同分类规则
X1	5	5	Y1
X2	12	12	Y1
X3	12	8	Y1
X4	3	3	Y4
X5	21	21	Y2
Y1	29	12	X2
Y2	21	21	X5
Y3	4	4	X3
Y4	3	3	X4
Sum	110	89	

通过表 1 可以计算得出规则集 X，Y 的正相似性评估为 80.90%。规则集 X，Y 正相似性计算就结束了，现在评估这两个规则集的负相似性，因为被误分的样本较少，故负相似性计算相对简单，从图 1 中可以看到规则集 X 误分 7 个样本，由规则 $X5$ 误分类，而规则

集 Y 误分 3 个样本,由规则 $Y2$ 误分。样本集中共有 7 个样本被误分,而被 X,Y 共同误分且结果一致的样本数为 3,这样可以得出两者负相似性达到了 42.85%,这也就意味着如果组合分类模型由这两个规则集组成,那么此组合分类模型分类精度与单模型分类精度相似,没有太多的提高。

2　实验验证

本实验使用 C4.5 决策树和 CART 决策树从训练样本中提取多个规则集,分别计算每类算法提取的规则集间的正相似性与负相似性,并以每类算法所建的模型为组合分类模型的基模型,分别建立两个组合分类模型,测试两个组合分类模型的精度来证明负相似性可以有效地选择基模型。实验的输入数据为应届毕业生信息数据,用下面规则对每个样本分配类标号:

IF ("某学生修完规定学分 = Yes" AND "毕业论文答辩合格 = Yes" AND "每门课平均 成绩≥80" AND "无违反校纪行为 = Yes") THEN class = 优秀应届毕业生;

IF ("某学生修完规定学分 = Yes" AND "毕业论文答辩合格 = Yes" AND "每门课平均成绩 < 80") THEN class = 普通应届毕业生;

IF ("某学生修完规定学分 = Yes" AND "毕业论文答辩合格 = Yes" AND "无违反校纪行为 = No") THEN class = 普通应届毕业生;

IF ("某学生修完规定学分 = No" OR "毕业论文答辩合格 = No") THEN class = 肄业生。

先建立 8 个数据集,每个包含 300 个样本,C4.5 算法用每个数据集训练一个 C4.5 决策树[5]。并建立一个测试样本集来测试这些决策树的性能,平均分类准确率为 96.84%。通过文中提出的相似性度量可以计算得出在不同数据集上提取的规则集间的正相似性与负相似性。测试了每个 C4.5 决策树与其它 C4.5 决策树之间的正相似性,结果如表 2 所示,平均正相似程度为 86.67%。

也对 CART 决策树算法进行了同样的操作,所生成的 8 个 CART 决策树,平均分类准确率为 97.25%,从分类准确率角度来看,CART 决策树与 C4.5 决策树两者大致相同,但是 CART 决策树算法所提取的规则集的正相似性得到了提高。实验结果见表 3,所有规则集的平均正相似程度达到了 90.21%。

通过规则集间正相似性的计算,认为 CART 决策树算法更适合这个数据集,虽然它生成的规则集与 C4.5 决策树算法生成的规则集分类精确度相似,但它所生成的规则集一致性更好。

表 2　C4.5 决策树之间的正相似性

模型	2	3	4	5	6	7	8
1	0.89	0.87	0.89	0.81	0.84	0.88	0.87
2		0.91	0.85	0.89	0.85	0.86	0.84
3			0.92	0.93	0.88	0.88	0.83
4				0.92	0.82	0.81	0.85
5					0.91	0.87	0.83
6						0.87	0.81
7							0.89

表 3　CART 决策树之间的正相似性

模型	2	3	4	5	6	7	8
1	0.91	0.89	0.91	0.93	0.91	0.88	0.91
2		0.87	0.88	0.91	0.89	0.91	0.92
3			0.89	0.87	0.92	0.89	0.89
4				0.92	0.91	0.86	0.87
5					0.93	0.92	0.93
6						0.94	0.88
7							0.92

同时也计算了规则集间的负相似性。测试了每个 C4.5 决策树与其它 C4.5 决策树之间的负相似性,结果如表 4 所示,平均负相似程度为 30.35%。而 CART 决策树之间的平均负相似性为 44.85%,如表 5 所示。现在,建立两个组合分类模型 A、B,其中 A 组合分类模型的基分类器为上面所建立的 8 个 C4.5 决策树,B 组合分类模型的基模型为上面所建立的 8 个 CART 决策树,组合分类模型对样本分类后采用大数表决的方式形成最终分类结果。用上面所建立的测试集进行测试,通过测试发现,A 组合分类模型的预测准确率为 97.86%,而 B 组合分类模型的预测准确率为 97.59%。这就说明了组合分类器的基模型之间负相似性越小越好,可以通过计算不同规则集之间的负相似性为组合分类器挑选基模型。

表 4　C4.5 决策树之间的的负相似性

模型	2	3	4	5	6	7	8
1	0.31	0.29	0.31	0.35	0.27	0.31	0.28
2		0.25	0.28	0.33	0.28	0.26	0.34
3			0.29	0.32	0.25	0.29	0.29
4				0.28	0.35	0.36	0.33
5					0.34	0.27	0.27
6						0.35	0.34
7							0.31

表 5　CART 决策树之间的的负相似性

模型	2	3	4	5	6	7	8
1	0.43	0.45	0.52	0.41	0.45	0.38	0.47
2		0.51	0.54	0.44	0.45	0.46	0.48
3			0.57	0.33	0.48	0.34	0.43
4				0.34	0.42	0.51	0.45
5					0.41	0.47	0.46
6						0.46	0.51
7							0.39

(下转第 79 页)

象;另外,特征的查询需要消耗很长的时间.从而还会使识别速度下降。在相同的测试集上 WINNOW 方法从训练到测试的时间为 22 分钟[7],SVM 方法的时间复杂度为 $O(n^3)$[5]。

分析三种方法可以发现:目前,语块识别存在如下两大缺点:

1)英语语块识别的策略是把语块识别问题转为类似词性标注的分类问题来解决,这种方法的缺点是无法顾及每个短语内部的组成特点。

2)传统的英语语块识别使用同一个模型和相同种类的特征。这种方法的局限性在于相同种类的特征无法同时适合多种短语类型,同时,数据稀疏现象也随之而来。

如果为了避免数据稀疏而只采用"词性"特征来识别多种语块,那些对于"词"敏感的短语准确率将会很低。因此针对不同的语块采用不同的特征和策略,不同短语的识别相互借鉴,最后把不同语块的识别集成在一起,将会起到很好的效果。

参考文献:

[1] Berwick R, Abney S, Tenny C. Parsing By Chunks: Principle - Based Parsing[M]. Dordrecht: Kluwer Academic Publishers, 1991.

[2] Abney S. Partial parsing via finite - state cascades[C]//Workshop on Robust Parsing. 8th European Summer School in Logic, Language and Information conference. Prague, Czech Republic: [s. n.], 1996: 8 - 15.

[3] Sang T K. Introduction to the CoNLL - 2000 Shared Task: Chunking[C]//Proceedings of CoNLL - 2000 and LLL - 2000 conference. Lisbon, Portugal: [s. n.], 2000: 127 - 132.

[4] Skut W, Brants T. A maximum - entropy partial parser for unrestricted text[C]//In Proceedings of the 6th Workshop on Very Large Corpora Conference. Montreal, Quebec: [s. n.], 1998.

[5] Kudoh T, Matsumoto Y. Use of Support Vector Learning for Chunk Identification[C]//Proceedings of CoNLL - 2000 and LLL - 2000 conference. Lisbon, Portugal: [s. n.], 2000: 127 - 132.

[6] Sang T K. Memory - Based Shallow Parsing[C]//In proceedings of CoNLL - 2000 and LLL - 2000 conference. Lisbon, Portugal: [s. n.], 2000: 559 - 594.

[7] Zhang T, Damerau F, Johnson D. Text Chunking based on a Generalization of Winnow[J]. Machine Learning Research, 2002, 2(2): 615 - 637.

[8] Zhao Jun, Huang ChangNing. The model of Chinese base noun phrase identification based transfer[J]. Journal of Chinese Information Processing, 1999, 13(2): 1 - 7.

[9] Zhang YiQi, Zhou Qiang. The auto identification of Chinese base phrase[J]. Journal of Chinese Information Processing, 2003, 16(3): 1 - 8.

[10] Li Heng, Zhu JingBo, Yao TianShun. The Chinese chunking using SVM[J]. Journal of Chinese Information Processing, 2004, 18(2): 1 - 7.

[11] Li SuJian, Liu Qun. The definition and establish of Chinese phrases[C]//JSCL - 2003 Conference. Harbin: [s. n.], 2003: 100 - 115.

[12] Littlestone N. Learning quickly when irrelevant attributes abound: a new liner - threshold algorithm[J]. Machine learning, 1988(2): 285 - 318.

[13] Zhang Tong, Damerau F, Johnson D. Text Chunking using Regularized Winnow[C]//In: Proceedings of ACL - 2001. Toulouse, France: [s. n.], 2001.

[14] Duda R O, Hart P E, Stork D G. Pattern Classification[M]. Beijing: China Machine Press, 2003.

(上接第 75 页)

3　结束语

文中所讨论的规则集的相似性度量,可以有效地选择模型和算法。提出了一个可以测量不同规则集正相似性与负相似性的算法,它可以灵活地应用在各种实际情况中。实验显示这种相似性度量方法可以帮助选择合适的算法以及组合分类模型中的基模型。

参考文献:

[1] Johansson U, Konig R, Niklasson L. Automatically balancing accuracy and comprehensibility in predictive modeling[C]// in: Proceedings of the 8th International Conferenceon Information Fusion. [s. l.]: [s. n.], 2005.

[2] Neumann J. Classifcation and evaluation of algorithms for rule extraction from artifcial neural networks[D]. summer project, University of Edingburgh, 1998.

[3] Lele S, Golden B, Ozga K, et al. Clustering rules using empirical similarity of support sets[C]// in: Proceedings of the 4th International Conferenceon Discovery Science. London, UK: Springer - Verlag, 2001: 447 - 451.

[4] Huysmans J, Baesens B, Vanthienen J. A new approach for measuring rule sets consistency[D]. Data & Knowledge Engineer, 2007, 63: 167 - 182.

[5] Quinlan J. C4. 5: Programs for Machine Learning[M]. San Francisco, CA, USA: Morgan Kaufman, 1993.