基于遗传算法和信息熵的文本分类规则抽取方法研究*

唐 华，曾碧卿
（华南师范大学南海校区计算机工程系，广东 佛山 528225）

摘 要：针对数据挖掘中的文本分类问题，提出了一种基于遗传算法和信息熵的文本分类规则抽取算法 GeneticMiner（简称 GM），该算法的目标是在数据集中发现分类规则。首先利用信息熵生成初始种群，然后利用优化的遗传算法抽取相应规则。采用六个标准的公共领域的数据集比较了 GM 与其它两个非常著名的同类算法 AntMiner和 CN2。实验结果表明，无论是预测准确性和规则的简单性，GM 都明显优于 AntMiner和 CN2。并且该算法能大大提高对知识的理解力。

关键词：文本分类规则；知识发现；信息熵；遗传算法；数据挖掘
中图分类号：TP182 文献标识码：A 文章编号：0529-6579（2007）05-0018-05

数据挖掘（Data Mining）已成为计算机科学中一个重要的研究方向，其目标是从数据中抽取知识。文本分类是数据挖掘研究的重点内容之一，具有广泛的实际应用[1-3]。目前常用的文本分类方法有贝叶斯分类方法、KNN 方法、支持向量机（SVM）和神经网络方法等[4-5]，这些方法追求的是较高的文本分类正确率，但不能抽取出使人易于理解的文本分类规则。现在虽然也有少数基于规则抽取的文本分类方法，但是这些分类方法抽取易于理解的分类规则仍然较困难。本文提出了一种基于信息熵的特征选取，然后用遗传算法进行文本分类、抽取分类规则的方法 GeneticMiner（简称 GM），以提高分类精度，优化分类规则，并提高对知识的理解力。

该算法的优化框架如图 1所示，执行流程可描述如下：

（1）变量初始化。将已发现规则列表设置为空，同时将所有的训练样本放置到训练样本集中。

（2）遗传算法的演化。遗传算法的每次演化都能发现一个分类规则。遗传算法演化完成后，将本次演化发现的规则加入到已发现规则列表中；同时，将该规则所覆盖的样本从训练样本集中剔除。

（3）终止条件。当未覆盖样本的数目小于用户预设值时，遗传算法停止演化。

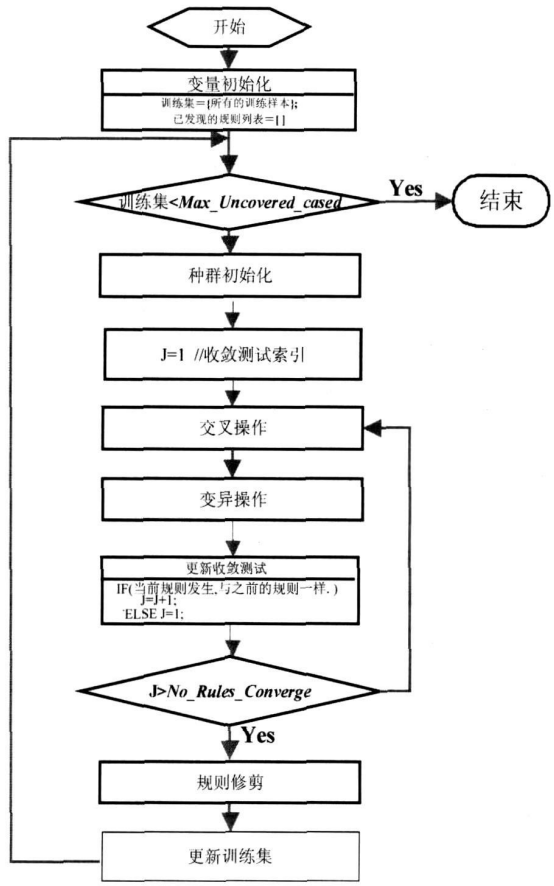

图 1 GM 的高层描述

Fig. 1 The high-level description of GeneticMiner

1 基于信息熵的初始种群生成方法

以下阐述基于信息熵的初始种群生成方法。

1.1 术语选择概率

设 $term_{ij}$表示条件 $A_i=V_{ij}$，A_i 代表第 i个属性，

* 收稿日期：2007-01-09
基金项目：国家自然科学基金资助项目（60573127）
作者简介：唐华（1973年生），男，讲师；Email：hua200@gmail.com

V_{ij}表示第个属性 A_i的第 j个取值；术语 $term_{ij}$被选择加入到当前规则中的概率为：

$$P_{ij}=\frac{x_i\cdot\eta_{ij}\cdot\tau_{ij}}{\sum_{i=1}^{\alpha}\left[x_i\cdot\sum_{j=1}^{b_i}(\eta_{ij}\cdot\tau_{ij})\right]} \quad (1)$$

其中，η_{ij}表示术语 $term_{ij}$的启发函数值，η_{ij}越大，术语 $term_{ij}$与当前分类的相关性越大，术语 $term_{ij}$以一个更高的概率被加入到当前的分类规则中；η_{ij}表示术语 $term_{ij}$被选择加入到分类规则中的次数，一些表现优秀的术语将会被越来越多地加入到当前的规则中，同时它们被选择加入到分类规则中的次数也越来越大；α表示属性的数目；x_i是一个布尔变量，如果 A_i还没有被选入到当前染色体中，则 x_i取值为 1，反之，取值为 0；b_i表示第 i个属性 A_i的可以取值的数目。

1.2　术语启发函数

对于每个能加入到当前分类规则中的术语 $term_{ij}$，GM通过一个启发函数来估算该术语的品质特征，即能提高当前规则预测准确性的能力。这个启发函数基于信息熵理论[6]构建而成。也就是说，η_{ij}的值反映了术语 $term_{ij}$的平均信息量的度量。对于每个术语 $term_{ij}$，它的信息熵可计算如下：

$$H(W\mid A_i=V_{ij})=-\sum_{w=1}^{k}(P(w\mid A_i=V_{ij})\cdot\log_2P(w\mid A_i=V_{ij})) \quad (2)$$

其中，W表示当前的训练样本集，k表示当前训练样本的数目，$P(w\mid A_i=V_{ij})$表示术语 $term_{ij}$出现在训练样本 w中的经验概率值。$H(w\mid A_i=V_{ij})$的值越大，表示属性 A_i的取值分布得越均匀，此时术语 $term_{ij}$将以较小的概率加入到当前的分类规则中。术语 $term_{ij}$的启发函数为：

$$\eta_{ij}=\frac{\log_2k-H(W\mid A_i=V_{ij})}{\sum_{i=1}^{a}x_i\cdot\sum_{j=1}^{b_i}(\log_2k-H(w\mid A_i=V_{ij}))} \quad (3)$$

1.3　术语插入条件

术语 $term_{ij}$在同时满足以下两个条件后，再以概率 p_{ij}加入到当前的分类规则中：

（1）属性 A_i还没被加入到当前的分类规则中；

（2）术语 $term_{ij}$加入到当前分类规则中后，该规则所覆盖的训练样本的数目必须大于用户预设的阈值（每个规则所覆盖的最小样本数目）。

1.4　术语插入终止条件

若下面的某个条件满足后，则停止向当前分类规则中添加术语：

（1）当前分类规则中术语的数目大于等于当前属性的数目；

（2）当前没有术语满足术语插入条件。

基于信息熵的初始种群生成方法的执行流程描述如下：

步骤一 生成一个空染色体；

步骤二 以术语选择概率随机选择一个术语；

步骤三 如果已选择的术语满足术语插入条件，则将术语插入到当前染色体中；

步骤四 如果当前的状况不满足术语插入终止条件，则转入步骤二；否则转入步骤五；

步骤五 为当前染色体选择一个结果术语（例如，预测的类名）。该结果术语要能最大程度地提升当前规则（当前染色体所对应的规则）的质量（该规则所覆盖的训练样本的数目）；

步骤六 重复步骤一至步骤五，直到成功产生所有的初始个体。

2　遗传算法的设计

遗传算法的设计包括染色体编码，适应度评价函数，交叉操作和变异操作等，以下分别予以阐明。

2.1　染色体编码

在遗传算法的演化过程中，那些适应度比较低的个体将会被逐渐淘汰，同时那些适应度比较高的个体将会被逐渐保留下来。目标函数取得最优值（最大值或最小值）的点就是优化问题的理想解。本文的优化问题就是挖掘出蕴含在训练样本中的分类规则，采用一个数组来表示对应于单个分类规则的一个染色体。对应于单个分类规则的染色体的编码方式如下：

t_1	t_2	…	t_n	c

这里，t_i表示分类规则中 $term_i$的取值，c表示分类规则中的取值。

2.2　适应度评价函数

笔者主要从支持度、置信度和覆盖度这三个方面来设计本文遗传算法的适应度评价函数。适应度函数的设计应综合考虑规则与样本的特征属性部分、类别属性部分及二者匹配情况。在进化的过程中各规则依靠自身适应度函数与其它规则进行竞争，只有支持度、置信度和覆盖度都高的规则，才能在竞争中生存下来。本文的适应度评价函数定义如下：

$$Fitness(r)=\alpha\times SI(r)+\beta\times CI(r)+\gamma\times\mathrm{II}(r) \quad (4)$$

其中，各个符号的含义解释如下：

N为待挖掘样本的数目；$R_c(r)$为满足给定规则条件部分的样本数目；$R_d(r)$为满足给定规则 r结果部分的样本数目；$R_e(r)$为满足给定规则 r的样本数目；$SI(r)$为给定规则 r的支持度，$SI(r) = R_e(r)/N$；$CI(r)$为给定规则 r的置信度，$CI(r) = R_e(r)/R_c(r)$；$\Pi_c(r)$为给定规则 γ的覆盖度，$\Pi(r) = R_e(r)/R_d(r)$；α, β, γ分别表示支持度、置信度和覆盖度的权值。

2.3 交叉操作

本文的交叉方式为在双亲（两个父辈的个体）的染色体上随机地选择一个断点，然后将断点上的基因互相交换，从而形成两个新的后代（图 2）。

传统的交叉操作，在所有基因位上随机地选择一个断点并进行相应操作，这种交叉操作模式的效率不高。如果这个断点选择到合适的位置上，那么交叉操作所产生的后代将得到改善；否则，交叉操作所产生的后代不能得到改善。为了增强本文交叉算法的效率，笔者设计了图 3所示的多次交叉取优的交叉模式。在该模式下，通过多次地随机地选取交叉断点，极大地提高了选择到合适断点位置的概率。这使得交叉操作所产生后代被改善的可能性得到了极大地增加。

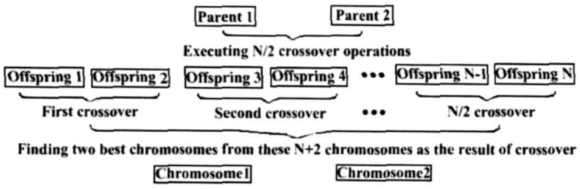

图 2 交叉操作

Fig. 2 The crossover operation of this paper

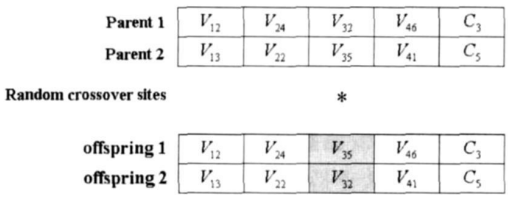

图 3 交叉方案的本质示意图

Fig. 3 Illustration of essential crossover scheme

2.4 变异操作

为了防止所有的解（染色体）都收敛于局部最优点，变异操作对当前染色体上的某个基因位上的基因进行变异。变异操作在染色体上自发地产生随机的变化，可以提供初始群体中没含有的基因或找回选择过程中丢失的基因，为群体提供新的内容。本文变异操作的一个范例如图 4所示。

为了提高本文变异算子的效率，笔者设计了图 5所示的多次变异取优值的变异模式。在该模式下，通过多次地随机地选取变异位置，极大地提高了选择到合适变异位置的概率。这使得变异操作所产生后代被改善的可能性得到了极大地增加。

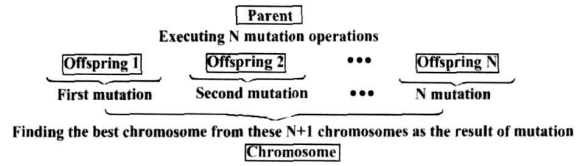

图 4 变异操作

Fig. 4 The mutation operation

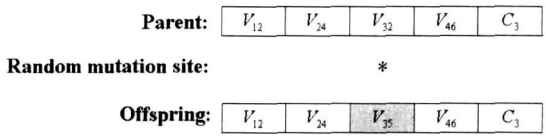

图 5 变异方案的本质示意图

Fig. 5 Illustration of essential mutation scheme

3 计算结果和讨论

采用六个标准的公共领域的数据集来测试本文 Genetic-Miner的绩效。首先，在预处理阶段将连续变量离散化。笔者采用 C4.5-Disc方法来完成变量的离散化工作[7]。该方法采用众所周知的 C4.5算法将连续变量进行离散化[8]。本文实验中的参数设置如下：

①*Min_cases_per_rule*=10; (user-specified threshold)
②*Max_uncovered_cases*=10; (user-specified threshold)
③*No_rules_converg*=10; (user-specified threshold)
④α=1.0; (weight of sustaining index)
⑤β=1.0; (weight of creditable index)
⑥γ=1.0; (weight of inclusive index)
⑦P_m=0.6; (mutation rate)
⑧P_x=0.8; (crossover rate)

本文方法的绩效与 CN2[9-10]和 AntMiner[11]进行了比较。实验中用到的数据集如表 1所示，三种方法预测的准确性的比较结果如表 2所示。笔者采用众所周知的 10-折交叉确认方法来度量分类结果的准确性[12]。可以看出，在对这六个数据集进行分类规则的抽取时，本文 Genetic-Miner方法抽取的规则的预测准确性较高。三种方法抽取规则的简单性如表 3所示。在这六个数据集中，采用本文 Genetic-Miner方法挖掘的规则最简单。综合考虑预测准确性和规则的简单性，本文方法明显优于 AntMiner和 CN2.

表 1　实验中用到的数据集

Tab. 1　Data Sets Used in the Experiments

Data set name	Number of cases	Number of Categorical Attributes	Number of Continuous Attributes	Number of Classes of the data set
Ljubljana breast cancer	282	9	—	2
Wisconsin breast cancer	683	—	9	2
Tic tac toe	958	9	—	2
Dermatology	366	33	1	6
Hepatitis	155	13	6	2
Cleveland heart disease	303	8	5	5

表 2　不同方法的预测精度

Tab. 2　Predictive Accuracy of Different Methods

Data set	GeneticMiner's Predictive Accuracy/%	Ant-Miner's Predictive Accuracy/%	CN2's Predictive Accuracy/%
Ljubljana breast cancer	81.36	75.28	67.69
Wisconsin breast cancer	98.15	96.04	94.88
tic tac toe	97.86	73.04	97.38
dermatology	95.61	94.29	90.38
Hepatitis	93.08	90.00	90.00
Cleveland heart disease	63.58	59.67	57.48

表 3　通过这些方法发现规则的简单性

Tab. 3　Simplicity of Rule Lists Discovered by These Methods

Data set	No. of rules (No. of terms/No. of rules)		
	GeneticMiner	Ant-Miner	CN2
Ljubljana breast cancer	6.89 (1.26)	7.10 (1.28)	55.40 (2.21)
Wisconsin breast cancer	6.13 (1.86)	6.20 (1.97)	18.60 (2.39)
tic tac toe	7.68(1.24)	8.50 (1.18)	39.70 (2.90)
dermatology	7.29 (2.39)	7.30 (3.16)	18.50 (2.47)
Hepatitis	3.28 (1.56)	3.40 (2.41)	7.20 (1.58)
Cleveland heart disease	8.64 (1.69)	9.50 (1.71)	42.40 2.79)

4　结　论

本文提出了一种有效的规则抽取方法 (GeneticMiner)。采用六个标准的公共领域的数据集来比较本文方法和其它两种方法的绩效。实验结果表明，无论是预测准确性和规则的简单性，本文方法都明显优于 Ant-Miner和 CN2.

参考文献：

[1] BOSE I, MAHAPATRA R K. Business data mining a machine learning perspective [J]. Information & Management, 2001, 39(3): 211 - 225.

[2] 王明春, 王正欧, 张楷, 等. 一种基于 CHI值特征选取的粗糙集文本分类规则抽取方法[J]. 计算机应用, 2005, 25(5): 1026 - 1028.

[3] SHI Y F, ZHAO Y P. Comparison of Text Categorization Algorithms[J]. 武汉大学学报: 自然科学英文版, 2004, 9(5): 798 - 804.

[4] TAN K C, YU Q, LEE T H. A distributed evolutionary classifier for knowledge discovery in data mining[J]. IEEE Transactions on Systems, Man and Cybernetics, Part C: Applications and Reviews, 2005, 35(2): 131 - 142.

[5] YANG YIMING. An evaluation of statistical approaches to text categorization[J]. Journal of Information Retrieval, 1999(1/2): 67 - 88.

[6] COVER T M, THOMAS J A. Elements of Information Theory[M]. New York: John Wiley Press, 1991.

[7] KOHAVI R, SAHAMI M. Error - based and entropy - based discretization of continuous features[C]. Proceedings of second international conference on Knowledge Discovery and Data Mining. Menlo Park, USA, 1996.

[8] QUINLAN J R. C4.5: Programs for Machine Learning [M]. San Francisco, CA: Morgan Kaufmann Publishers Inc, 1993.

[9] CLARK P, NIBLETT T. The CN2 induction algorithm [J]. Machine Learning, 1989, 3(4): 261 - 283.

[10] CLARK P, BOSWELL R. Rule induction with CN2: Some recent improvements[C]. Lecture Notes in Artificial Intelligence. Berlin: Springer - Verlag, 1991: 151 - 163.

[11] PARPINELLI R S, LOPES H S, FREITAS A. A Data Mining With an Ant Colony Optimization Algorithm[J]. IEEE Transactions on evolutionary computing, 2002, 6(4): 321 - 332.

[12] WEISS S M, KULIKOWSKI C A. Computer Systems that Learn[M]. San Francisco, CA: Morgan Kaufmann Publishers Inc, 1991.

(下转第 24页)

$v_q) \notin E(G)(p \neq q, p, q \neq 1)$即图 G同构于 G_2。

情况 2 $[V(G) N(D)] = \varnothing$

此时：$|V(G)| = |V(D)| = n$，而 $\gamma(G) = 2$ 且 $\gamma(G) = l(D) - 1 = n - 1$，则 $n=3$，即 G同构于一个三角形。

定理 1、2和 3给出了 n阶本原无向图 G中指标 $k(G)$的最大值和次大值，并且刻画了达到最大值和次大值的所有 n阶本原无向图。

参考文献：

[1] 邵嘉裕. 对称本原矩阵的指数集[J]. 中国科学：A辑，1986(9)：931－939.

[2] LIU B, MECKY B D, WORMALD N C, et al. The exponent set of symmetric primitive (0, 1) matrices with zero trace[J]. Linear Algebra Appl, 1990(133)：121－131.

[3] LIU B, SHAO J Y. Characterization of primitive extremal matrices[J]. 中国科学：A辑(英文版)，1991(7)：25－36.

[4] CAI J, WANG B. The characterization of symmetric primitive matrices with exponent n－1[J]. Linear Algebra Appl, 2003(364)：135－145.

[5] CAI J, ZHANG K. The characterization of symmetric primitive matrices with exponent 2n－2r[J]. Linear Multilinear Algebra, 1995(39)：391－396.

[6] BRUALDI R A, RYSER H J. Combinatorial matrix theory[M]. Cambridge：Cambridge University Press, 1991.

[7] 王建中，王殿军. 对称本原矩阵的指数集的刻划[J]. 数学进展，1994，22(5)：516－523.

[8] WEI F, CHU M, WANG J. On odd primitive graphs [J]. Australas J Combin, 1999(3)：11－15.

[9] ZHOU B. Exponents of primitive graphs[J]. Australas Combin, 2003(28)：67－72.

[10] BRUALDI R A, LIU Bo－lian. Generalized exponents of primitive directed graphs[J]. Graph Theory, 1990(14)：483－499.

[11] LIU Bo－lian. Generalized exponents of boolean matrices[J]. Linear Algebra and its Applications, 2003(373)：169－182.

Primitive Graphs of Extremal Local Exponent

WEI Fu-yi, ZENG Wen-cai, LIU Mu-huo

(Department of Applied Mathematics, South China Agricultural University, Guangzhou 510640, China)

Abstract: Let G be the primitive graphs of order n, and $k(G)$ be the total number of local exponent which is equal to the exponent of G. All the primitive graphs of order n for which $k(G)$ achieve the largest and the second largest value are characterized completely.

Key words: exponents; primitive graphs; associated matrix; local exponent

(上接第 21页)

Research on Method of Text Classification Rule Extraction Based on Genetic Algorithm and Entropy

TANG Hua, ZENG Biqing

(Computer Engineering Department of Nanhai Campus, South China Normal University, Foshan 528225, China)

Abstract: Aimed at the text classification problems in data mining, a text classification rule extraction method is proposed based on genetic algorithm and entropy for rule discovery called Genetic-Miner (GM). The goal of GM is to discover classification rules in data sets. It produces population with the entropy and then extract classification rule with genetic algorithm. Compared the performance of GM with other two well-known algorithms Ant-miner and CN2 in six public domain data sets, the results showed that GM has a better performance in both predictive accuracy and rule list simplicity criteria than Ant-Miner and CN2. It can also mostly improve the comprehensibility of the discovered knowledge.

Key words: text classification rule; data mining; discover knowledge; information entropy; genetic algorithm