文章编号:1001 – 893X(2010)05 – 0023 – 05

基于粗糙集的 CBR 系统案例检索策略*

孙岩清[1,2],尹树华[3],林初善[3]

(1.西安通信学院 研究生管理大队,西安 710106;2.中国酒泉卫星发射中心 指挥通信室,甘肃 酒泉 732750;
3.西安通信学院 军用光纤通信教研室,西安 710106)

摘 要:针对案例推理系统中案例检索的效率和质量问题,提出一种新的案例检索策略。采用粗糙集进行案例属性约简,完成案例库优化,并计算反映专家经验的属性权重,结合相似度计算和人工神经网络进行不同情况下的案例检索。运用 UCI 数据集进行了仿真对比,将其用于数字数据网故障诊断系统中,结果表明所提出的策略在不同数据集下均具有较高的检索效率,更加适用于实际 CBR 系统。

关键词:基于案例推理;概率神经网络;粗糙集;案例检索;故障诊断系统

中图分类号:TP391.3 **文献标识码**:A doi:10.3969/j.issn.1001 – 893x.2010.05.005

A Case Retrieval Strategy for CBR System Based on Rough Set

SUN Yan-qing[1,2], *YIN Shu-hua*[3], *LIN Chu-shan*[3]

(1. Department of Graduate Student Management, Xi′an Communication Institute, Xi′an 710106, China;
2. Command and Communication Room, Jiuquan Satellite Launch Centre of China, Jiuquan 732750, China;
3. Teaching Section of Military Optical Fiber Communication, Xi′an Communication Institute, Xi′an 710106, China)

Abstract: A new case retrieval strategy is proposed for case-based reasoning(CBR) system because of the case retrieval efficiency and quality problem. The rough set theory is adopted to implement case attribute reduction, complete the case base optimization, and compute attribute weights that reflect the expert′s experience firstly, and then is combined with similarity computation and artificial neural network(ANN) to accomplish case retrieval in different situation. The UCI data set is used to simulate and compare. Application of the retrieval strategy in the data digital network fault diagnosis system indicates that the proposed case retrieval strategy has better performance in different data sets, and it is more fit for practical CBR system.

Key words: case-based reasoning(CBR); probabilistic neural network; rough set; case retrieve; fault diagnosis system

1 引 言

基于案例推理(Case-Based Reasoning, CBR)是通过回忆一个或几个过去发生的具体案例,进而采用类比的推理方法,提出解决新问题的方案,其一般过程为“检索 – 重用 – 修正 – 存储”,检索是其中的关键,直接决定了案例推理系统的性能。目前,研究较多的检索方法有决策树[1]、KNN[2-3]、神经网络[4-5]、支持向量机[6]等,但其每一种具体算法都有一定的局限性,不能够在 CBR 系统中得到很好的应用。其中,决策树法存在案例库改变时需要重新建树且存储、开销大的缺点;神经网络法存在案例属性较多时训练耗时,只能给出单个相似案例的缺点;KNN 算法存在计算量大、效率不高和在案例较多时检索耗时的缺点;支持向量机则存在随着案例或案例属性增加而检索耗时、计算复杂的缺点。

因此,已有检索方法存在各自问题,不能很好地应用于实际的 CBR 系统,故本文提出基于粗糙集理论进行属性约简,删除案例冗余属性,完成案例库优

* 收稿日期:2010 – 02 – 08;修回日期:2010 – 03 – 23

化,再结合相似度计算方法和概率神经网络算法进行不同情况下的案例检索策略,做到既保证检索的精度,尽可能地检索出要求的所有相似案例,又避免检索时间随案例增加而线性增长。

2 粗糙集相关概念

2.1 属性重要度定义

定义 1:设 $S=(U,A,V,f)$ 为一个信息系统,$A=C\cup D$,$\forall R\subseteq C$,属性依赖度表示为 $r(R,D)=|Pos_R(D)|/|Pos_C(D)|$,则 $\forall c\in R$ 的属性重要度可表示为依赖度的差值:

$$SIG_R^1(c)=r(R\cup c,D)-r(R,D) \quad (1)$$

定义 2:设 $S=(U,A,V,f)$ 为一个信息系统,$A=C\cup D$,$\forall R\subseteq C$,且 R 在对象集合 U 上产生的划分为:$U/R=\{X_1,X_2,\cdots,X_n\}$,则知识 P 的熵为

$$H(R)=-\sum_{i=1}^{n}p(X_i)\,\mathrm{lb}\,(p(X_i))$$

式中,$p(X_i)=|X_i|/|U|$。

则决策表中任一条件属性本身的重要度可以由它所引起的信息熵的变化来衡量,即已知属性集 $R\subseteq C$,$\forall c\in C-R$ 的重要度可定义为

$$SIG^2(c,R,C)=H(R\cup c)-H(R) \quad (2)$$

对于 CBR 系统,约简应既能很好地反映专家经验知识,又能生成正确的决策规则,因此,应该综合考虑属性决策分类和本身重要度两方面的因素。

定义 3:对于决策信息系统 $S=(U,A,V,f)$,$A=C\cup D$,$n=|U|$,属性 $c\in R\subseteq C$ 在 R 中的重要度为

$$SIG(c)=SIG_R^1(c)+wSIG^2(c,R,C)/\mathrm{lb}(n) \quad (3)$$

式中,$0\leqslant w\leqslant 1$。当 $w=1$ 时,同等考虑属性对决策分类的影响度和属性本身的重要度,最大化地反映领域专家的经验知识;当 $w=0$ 时,仅考虑属性对决策分类的影响,而一般情况下,对于 CBR 系统采取前者的定义。

2.2 知识约简定义

定义 4:设 $S=(U,A,V,f)$ 为一个信息系统,$A=C\cup D$,$\forall P\subseteq C$,如果 P 满足下面两个条件,则 P 是一个 Pawlak 约简:

(1) $Pos_P(D)=Pos_C(D)$;

(2) $\forall a\in P, Pos_{P-\{a\}}(D)\neq Pos_C(D)$。

上面定义中,第一个条件保证了相同决策规则的生成,第二个条件保证了约简的独立性。

3 相似案例检索思想

3.1 案例相似度定义及分析

设 F 为一案例库,且其中案例的属性均已进行归一化处理。

定义 5:以 $dist(A,B)$、$sim(A,B)$ 分别表示案例 A、B 之间的距离和相似度,则在最近邻实例检索中 $sim(A,B)=1-dist(A,B)$,那么,$sim(A,B)$ 应满足以下条件和性质:

(1) $sim(A,B)\in[0,1]$,$sim(A,B)=1$,当且仅当 $A=B$,即自反性;

(2) $sim(A,B)=sim(B,A)$,即对称性;

(3)对任意的案例 $A,B,C\subset F$,有 $sim(A,B)\geqslant sim(A,C)+sim(B,C)-1$,即满足三角不等式关系。

由定义 5 可知,采用最近邻进行检索案例的核心工作就是计算目标案例与待检案例之间的距离,而后选取距离最小者作为最相似案例。在实际应用中多采用欧几里得距离法,同时,为满足条件(1),对传统距离公式进行改进,对距离进行归一化处理,有:

$$similarity(A,B)=1-\sqrt{\sum_{i=1}^{n}w_i\,(A(i)-B(i))^2/n} \quad (4)$$

式中,w_i 为案例的第 i 个属性权值,可以在属性约简的过程中获得,其值越大则表示该属性越重要;n 为属性个数;$A(i)$、$B(i)$ 分别表示案例 A、B 的第 i 个属性值。

3.2 案例检索过程

图 1 为案例检索流程图。

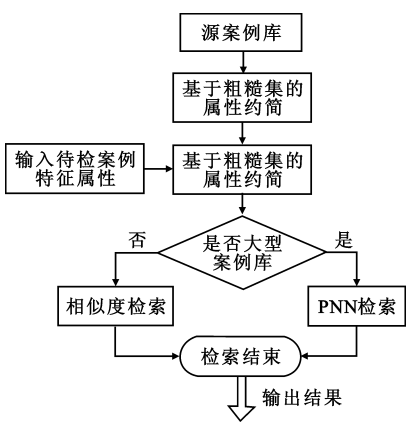

图 1 案例检索流程图

Fig.1 Case retrieval flowchart

利用粗糙集理论首先对案例库进行属性约简,

并计算约简后的属性重要度权值,而后在小案例库时采取相似度计算方法检索案例,在大案例库时采用概率神经网络实现,从而充分利用相似度计算和神经网络的优点,取长补短,达到 CBR 系统案例检索的最优效果。

4　实验结果和分析

为验证文中检索策略的正确性,采用 UCI 数据集和人工数据集相结合的方法进行,仿真环境为 Matlab R2006a,计算机配置为 AMD Athlon 64 位处理器,1G 内存。其中,UCI 数据集主要采用了 Wine、Riply 和 Iris 3 种,分别用于验证时间复杂度和检索精度,同时在小数据集下运用人工数据集对检索精度进行了验证。

4.1　案例检索时间复杂度验证

采用 Wine 数据集进行时间复杂度验证,它包括 178 个样本、13 个条件属性和 3 个决策属性。实验以成倍增加案例的方式进行,任选其中的一个案例作为待检测样本,同时,为避免检索时间的随机性,降低仿真误差,采取每次检索仿真 10 次,取平均值作为最终检索时间的方法。仿真结果如图 2 所示。

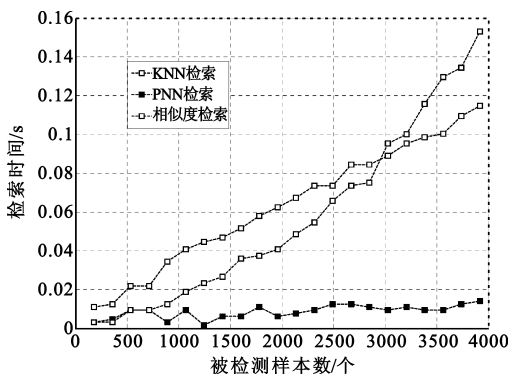

图 2　3 种检索方法的时间对比

Fig.2 The time comparison of three retrieval methods

由图 2 可以看出,在小数据集时,3 种检索算法耗时均很小,且相似度计算方法性能更优;而随着案例的增多,基于相似度计算和 KNN 算法的检索时间会线性增长,神经网络算法则在一定的时间点或范围内保持稳态。

4.2　案例检索精度验证

采用 Riply 数据集进行检索精度的验证,Riply 数据集包括训练样本 250 个、检测样本1 000个、条件属性 2 个、决策属性 2 个。检索结果如表 1 所示,其中相似度检索选择了两种模式,即取一个相似案例和两个相似案例。

表 1　3 种算法检索结果对比

Table 1 The retrieval result comparison of three algorithms

检索算法	检索精度/(%)
KNN($k=3$)	86.6
PNN	89.6
相似度检索	85.0(取一个案例) 92.2(取两个案例)

由表 1 可知,在只追求单个最相似案例的情况下,概率神经网络检索更加精确,K 近邻次之,相似度检索算法较差。但前两种算法却不能够给出多个相似案例,存在局限;而相似度检索算法则能够给出多个相似案例,一般选择 2 个,在此情况下,相似度检索算法具有相当高的精度,优势比较突出。

4.3　基于粗糙集的案例检索验证及应用

由以上实验可以看出:在小数据集时,相似度计算检索既能保证检索精度,又能保证检索的时间复杂度;在大数据集时,神经网络算法则可以保证检索精度,且能够避免检索时间的线性增长。因此,文中提出的案例检索策略能够有效提高 CBR 系统的性能,适合于案例推理的实际应用,结合粗糙集理论则能够进一步优化检索的时间复杂度问题。

用 Iris 数据集进行实验,它包括 150 个案例样本、4 个条件属性和 3 个决策属性,用其中 90 个样本进行训练,其余 60 个样本用于测试。运用 Matlab 对 3 种算法进行仿真,检索时间采取 10 次仿真的加权平均值,约简后训练数据集包含 88 个样本、3 个条件属性,属性重要度值分别为1.071 1、0.755 7 和 1.602 1。检索结果如表 2 所示。

表 2　约简前后的检索结果对比

Table 2 The retrieval result comparison of before-and-after reduction

检索算法	时间/s		精度/个	
	未约简	约简后	未约简	约简后
KNN($k=3$)	0.010 94	0.007 82	59	59
PNN	0.010 92	0.007 80	58	59
相似度检索	0.018 74	0.014 05	58	59

由表 2 可以看出,经过粗糙集约简后的案例检索算法,在案例检索效率和精度方面都有一定提高,尤其对于相似度检索方法,效果更加明显。由此可

以看出,利用粗糙集方法对案例库优化能够有效提高案例推理系统的检索效率,从而能够提高 CBR 系统的整体性能。

将基于粗糙集的案例检索策略应用于数字数据网故障诊断系统中,收集了网络运行中出现的 46 个典型案例,包括 9 个条件属性和 9 个决策属性,限于篇幅,具体含义不作详述。其中 38 个案例用于训练、8 个用于测试,分别如表 3 和表 4 所示。

表 3 训练案例表

Table 3 The training case table

案例	条件属性									决策属性
	a	*b*	*c*	*d*	*e*	*f*	*g*	*h*	*i*	
1	0	1	1	1	0	1	1	1	0	1
2	0	1	1	1	1	1	1	1	0	2
⋮				⋮						⋮
8	0.75	1	1	0	0	1	1	0	0	8
⋮				⋮						⋮
16	0.75	1	1	0	0	1	1	0	0	7
17	0	1	1	0	0	1	1	0	0	7
18	0.75	1	1	1	1	1	1	0	1	3
19	0	1	0	0	0	0	0	0	0	9
⋮				⋮						⋮
35	0.75	1	1	1	1	1	1	0	1	3
36	0	1	1	0	0	1	1	0	0	7
37	0.75	1	1	1	1	1	1	0	1	3
38	0	1	0	0	0	0	0	0	0	9

表 4 测试案例表

Table 4 The testing case table

案例	条件属性									决策属性
	a	*b*	*c*	*d*	*e*	*f*	*g*	*h*	*i*	
1	0	1	1	0	0	1	1	0	0	7
2	0	1	1	1	1	1	1	1	1	2
3	0.5	1	1	1	1	1	1	0	1	3
4	0.5	1	1	1	1	1	1	1	1	3
5	0	1	1	0	0	1	1	1	0	1
6	0.75	1	1	1	1	1	1	0	1	3
7	0.75	1	1	1	1	1	1	1	1	3
8	0	1	0	0	0	0	0	0	0	9

显然,表 3 中案例 8 和案例 16 为噪声案例,案例 36、37、38 为冗余案例。运用粗糙集进行属性约简,得到约简后的决策表,即删除了相同冗余案例 37、38,合并噪声案例 8 和 16 成一个新案例,约去了冗余属性 *c*。

由于案例库较小,采用相似度检索算法实现。约简后各属性重要度如表 5 所示,可以看出属性“*a*”和“*g*”的重要度明显大于其它属性的重要度,而它们分别代表终端数据收发情况和信道连接情况,对于信道类故障,它们也正是故障案例的重要特征,是专家判断故障类型的主要依据。可见,基于粗糙集的属性重要度值能真实反映属性的重要程度及专家经验。

表 5 基于粗糙集的属性重要度表

Table 5 The table of attribute significance based on rough set

	约简后条件属性							
	a	*b*	*d*	*e*	*f*	*g*	*h*	*i*
重要度	0.222 7	0.122 3	0.171 8	0.171 8	0.063 4	0.515 0	0.184 1	0.178 5

检索结果如表 6 所示,“/”两端分别表示基于粗糙集的属性重要度和默认属性重要度检索结果。当取相似案例数为 1 时,能够得到绝大部分待检案例的正确故障类别;当取相似案例数为 2 时,基于粗糙集重要度的相似度检索得到了所有正确类别,而基于一般默认属性重要度的相似度检索则仍不能涵盖所有的正确类别;当取数为 3 时,两种情况均涵盖了所有的正确类别。

因此,在实际应用中,相似度检索方法在案例库较小时能够尽可能地检索到所有相似案例,用于指导实际的故障诊断,而采用粗糙集重要度则能够进一步提高案例检索准确度,相对于一般默认属性重要度都为 1 的情况,案例的检索效率更高,也更有利于提高故障诊断的准确性。

表 6 粗糙集与默认属性重要度的相似度检索结果

Table 6 The similarity retrieval result of rough set and default attribute significance

	待检案例所属类别							
	7	2	3	3	1	3	3	9
所检最相似类别	7/7	2/2	3/3	3/3	4/4	3/3	3/3	9/9
所检次相似类别	8/8	2/2	2/2	3/3	1/6	2/2	3/3	8/4
所检较次相似类别	7/4	3/3	3/3	3/3	1/1	3/3	3/3	8/8

5 结 论

根据案例推理系统的实际,分析了反映专家经

验的属性重要度,结合粗糙集理论,提出了不同案例库下的案例检索方法,十分适用于增长式的案例推理系统。与前人单纯检索策略相比,文中充分利用粗糙集理论、相似度计算和神经网络等方法的各自优点,保证了 CBR 系统案例检索的精度和时间效率。实验结果表明,检索策略能够有效避免神经网络方法小案例库的精度较低和大案例库时相似度计算及 KNN 算法检索时间线性增长的缺点,将其应用于数字数据网故障诊断中,可以显著提高案例检索的精度,降低检索时间。但此检索策略不适用于动态案例库的情况,这方面的工作需要进一步研究。

参考文献:

[1] 王波,宋东,姜华男.基于粗糙集的 CBR 故障诊断案例的检索方法研究[J].计算机测量与控制,2007,15(11):1430 – 1433.
WANG Bo,SONG Dong,JIANG Hua-nan.Case Retrieve of Fault Diagnosis Expert System Based on CBR[J].Computer Measurement & Control,2007,15(11):1430 – 1433.(in Chinese)

[2] LI Yan, Simon C K Shiu, Sankar K Pal. Combining Feature Reduction and Case Selection in Building CBR Classifiers [J]. IEEE Transactions on Knowledge and data Engineering, 2006,18(3):415 – 429.

[3] 蒋占四,陈立平,罗年猛.最近邻实例检索相似度分析[J].计算机集成制造系统,2007,13(6):1165 – 1168.
JIANG Zhan-si,CHEN Li-ping,LUO Nian-meng.Similarity analysis in nearest-neighbor case retrieval[J].Computer Integrated Manufacturing Systms,2007,13(6):1165 – 1168.(in Chinese)

[4] Piliouras N, Kalatzis I, Theocharakis P. Development of the probabilistic neural network-cubic least squares mapping classifier to assess carotid plaques risk[J]. Pattern Recognition Letters,2004,25(2):249 – 258.

[5] WU Jian-da, CHIANG Peng-hsin, CHANG Yo-wei. An expert system for fault diagnosis in internal combustion engines using probability neural network[J]. Expert Systems with Applications,2008,34(4):2704 – 2713.

[6] 刘江永,王大明.基于支持向量机的快速高光谱分类研究[J].陕西师范大学学报(自然科学版),2009,37(4):43 – 47.
LIU Jiang-yong, WANG Da-ming. Fast classification of hyperspectral data based on support vector machines[J]. Journal of Shaanxi Normal University(Natural Science Edition),2009,37(4):43 – 47.(in Chinese)

作者简介:

孙岩清(1983 –),男,河南鹿邑人,硕士研究生,主要研究方向:Rough 集理论及应用、专家系统;

SUN Yan-qing(male) was born in Luyi, Henan Province, in 1983. He is now a graduate student. His research interests include rough set theory and its application and expert system.

Email:asdf_000009@163.com

尹树华(1952 –),男,陕西西安人,教授,主研研究方向:光通信技术应用、专家系统;

YIN Shu-hua(male) was born in Xi′an, Shaanxi Province, in 1952. He is now a professor. His research interests include optical communication technique application and expert system.

林初善(1979 –),男,福建福州人,硕士,助教,主要研究方向:光通信技术应用。

LIN Chu-shan(male) was born in Fuzhou, Fujian Province, in 1979. He is now a teaching assistant with the M.S. degree. His research direction is optical communication technique application.