ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 30,No. 6,Nov. 2015,pp. 1132－1340
DOI:10.16337/j.1004-9037.2015.06.024

http://sjcj.nuaa.edu.cn
E-mail:sjcj@nuaa.edu.cn
Tel/Fax:＋86-025-84892742

基于经典粗糙集的近似集动态获取方法

胡成祥　赵瑞斌

(滁州学院计算机与信息工程学院,滁州,239000)

摘　要:信息系统中的数据是动态变化的,根据动态变化的信息系统获取有用的信息,成为数据处理中的关键问题。针对该问题,分别讨论了信息系统中属性增加和减少时,近似集的动态获取方法。通过对信息系统中原有的等价类进行划分,避免了对论域的重新划分,提高了动态更新近似集的效率,通过讨论等价类与原有近似集之间的关系,给出了信息系统动获取之后的近似集与原来近似集之间的相关定理,提出了在经典粗糙集模型中,属性增减时近似集动态获取方法。实验结果验证了该方法的正确性和有效性,而且效率优于原始的方法。

关键词:知识发现;近似集;动态更新

中图分类号:TP18　　**文献标志码**:A

Approach for Dynamical Approximations Acquisition Based on Rough Set

Hu Chengxiang, Zhao Ruibin

(School of Computer and Information Engineering, Chuzhou University, Chuzhou, 239000, China)

Abstract: The data in information system is dynamically changed. How to acquisite useful information according to dynamical varied information system is a key problem in data processing. To deal with the problem, the approaches for dynamical approximations acquisition while adding or deleting an attribute are respectively discussed in information system. By dividing original equivalent classes in information systems, an approach which avoids re-division of the universe is proposed. The efficiency of dynamical updating approximation is improved. By analyzing the relationship between equivalent classes and original approximations, the corresponding theorems between updated approximations and original approximations are given. Then, the approaches for dynamical acquisition of approximations while adding or deleting an attribute are respectively proposed in classical rough set model. Experimental results verify the validity of the approaches and prove that the efficiencies of the proposed approaches are better than those of the original approaches.

Key words: knowledge discovery; approximation; dynamical updating

引　　言

粗糙集理论是由波兰数学家 Pawlak 提出的处理不确定性和模糊知识的数学工具[1],已经在专家系统、知

基金项目:国家自然科学基金(60873108)资助项目;安徽省高校自然科学基金重点(KJ2014A184)资助项目。
收稿日期:2014-06-05;**修订日期**:2014-06-30

识发现、图像图像处理、预测分析以及模式识别等方面有着广泛的应用[1-4],文献[5]提出了一种利用 SVM 和粗糙集理论相结合的方法对航空发电机诊断,文献[6]提出了利用粗糙集理论对图像进行增强的方法。

在现实生活中,各个行业的数据,随时随地都在发生着变化,使得大部分存储在数据库中的数据在不断地改变。当信息系统中的数据发生变化时,如果利用粗糙集理论的传统方法更新所需要的知识,会浪费大量时间进行重复计算,而且效率不高。因此,人们越来越重视研究怎样才能够对动态变化的信息系统中潜在的信息进行有效的更新。对于粗糙集理论方面,对动态知识获取的研究工作主要集中在不同的粗糙集模型中对象增减和属性增减时知识的动态获取。

Chan 在经典粗糙集模型中讨论了属性的上下边界域和等价类对近似集所产生的影响,研究了近似集和规则获取方法[7],这种方法可以通过计算信息系统中每个对象对应单个属性的边界域,只是所提出的算法时间复杂度较高。Li 等采用增量式更新近似集的方法,可以根据已有的近似集进行动态更新,提高了近似集更新的效率[8]。文献[9]以矩阵作为运算方法研究了当论域随时间变化时,变精度粗糙集模型中上、下近似集的增量式更新方法,并提出了一种基于矩阵的近似集增量式更新算法。文献[10,11]讨论了原近似集与已知单个属性的上边界域、下边界域之间的关系,研究近似集的获取方法,并讨论了在特性关系粗糙集模型中,近似集增量式获取方法,但没有分析算法时间的复杂度等,文献[12]在模糊信息系统中分析了对象集发生变化时近似集的更新理论,提出了一种基于高斯核模糊粗糙集模型的近似集获取方法,文献[13]讨论了动态环境下变精度粗糙集模型中,对象集发生变化时,信息粒度和近似集的变化情况,并且采用实验证明了方法的有效性,文献[14]提出了快速计算粗糙模糊集模型中概念的上近似集、下近似集的增量式方法,文献[15,16]用矩阵研究了集值信息系统中的属性集变化时概念近似集更新的方法,并且给出了一种计算近似集的并行化方法,文献[17]在不完备信息系统中,分析了动态更新近似集的方法。本文主要通过讨论经典粗糙集模型在信息系统中属性增加或者减少时等价类的变化情况,分析动态更新前和动态更新后近似集之间的关系,给出了动态获取上近似集和下近似集的相关定理,并在经典粗糙集模型中提出了属性增加或者减少时近似集动态获取算法。

定义 1[1] 设知识表达系统 $S=(U,A,V,f)$,其中 U 表示对象的非空有限集合,称为论域;A 表示非空属性的集合;V 表示属性值域的集合;$f:U\times A\to V$ 表示信息函数,为对象在属性上赋予的信息值,即 $\forall a\in A, x\in U, f(x,a)\in V_a$。若存在 $x\in U, a\in C, f(x,a)$ 未知($f(x,a)=*$),则称该知识表达系统是不完备的;否则称该知识表达系统是完备的。

定义 2[1] 对于属性子集 $B\subseteq A$,定义不可分辨二元关系 $\mathrm{IND}(B)$,即

$$\mathrm{IND}(B)=\{(x,y)\mid (x,y)\in U^2, \forall b\in B(b(x)=b(y))\} \tag{1}$$

显然,$\mathrm{IND}(B)$是一个等价关系,且 $\mathrm{IND}(B)=\bigcap\limits_{b\in B}\mathrm{IND}(\{b\})$

定义 3[1] 信息系统 $S=(U,A,V,f)$,对任意子集 $X\subseteq U$ 和其中的等价关系 $R\in\mathrm{IND}(B)$,X 的 R 上近似集和和下近似集分别定义为

$$\underline{R}(X)=\bigcup\{Y\in U/R\mid Y\subseteq X\} \tag{2}$$

$$\overline{R}(X)=\bigcup\{Y\in U/R\mid Y\cap X\neq\varnothing\} \tag{3}$$

它们的等价形式可表示为

$$\underline{R}(X)=\{x\in U\mid [x]_R\subseteq X\}, \overline{R}(X)=\{x\in U:[x]_R\cap X\neq\varnothing\} \tag{4}$$

1 属性增加时近似集动态获取方法

当信息系统因属性增加而发生动态变化时,传统的解决方法是对变化之后的信息系统进行重新计算等价类、近似集和决策规则。这样做需要对信息系统进行重新计算,效率比较低。因此,在单个属性增加时,通过增加的属性对原信息系统中已知的等价类进行重新划分,根据划分结果重新计算上近似集和下近似集,这样可以减少由于重复计算所花费的时间,提高效率。

1.1 增加单个属性时近似集获取方法

在已知的信息系统中,当单个属性增加时,原有的等价类会因为增加属性而发生变化,从而被划分的更细,等价类的数量会随之增加,因此上近似集有减小的趋势,下近似集有增大的趋势。下面给出增加单个属性时近似集获取的相关定理。

定义 4 设信息系统 $S=(U,A,V,f)$,对论域 U 进行划分形成的等价类为 $U/R=\{E_1,E_2,\cdots,E_n\}$,属性增加前集合 X 的 R 上近似集和 R 下近似集分别记为 $\overline{R}(X)$ 和 $\underline{R}(X)$。当属性增加后,如果 $f_{\exists x_i\in E_i}(d)\neq f_{\exists x_j\in E_i}(d)(i,j=1,2,\cdots,n,i\neq,j)$,那么等价类 $E_i(i=1,2,\cdots,n)$ 被细化。等价类 E_i 被划分更细之后形成新的等价类为 $E_{ik}(k=1,2,\cdots,j_i)$,属性增加后集合 X 的 R 上近似集和 R 下近似集分别记为 $\overline{R^{\vee}}(X)$ 和 $\underline{R^{\vee}}(X)$。

推论 1 设信息系统 $S=(U,A,V,f)$,对论域 U 进行划分的等价类 $E_i(i=1,2,\cdots,n)$,信息系统中增加的属性为 d,信息函数为 f。

(1)若 $f_{\forall x_i\in E_i}(d)=f_{\forall x_j\in E_i}(d)(i,j=1,2,\cdots,n,i\neq j)$,则属性增加后,等价类 E_i 不再被划分得更细,此时,$\underline{R^{\vee}}(X)=\underline{R}(X)$,$\overline{R^{\vee}}(X)=\overline{R}(X)$。

(2)属性增加后,若等价类被划分的更细,等价类 $E_i(i=1,2,\cdots,n)$ 细化所形成的新的等价类记为 $E_{ik}(k=1,2,\cdots j_i)$,如果 $E_i\subseteq\underline{R}(X)$,那么 $E_{ik}\subseteq\underline{R^{\vee}}(X)$。

(3)属性增加后,若每个对象对应属性 d 的属性值都不相等,那么 $\underline{R^{\vee}}(X)=\overline{R^{\vee}}(X)=X$。

(4)属性增加前,若等价类 E_i 与集合 X 无交集,即 $E_i\cap X=\varnothing$,则属性增加后,近似集不变。

定理 1 给定信息系统 $S=(U,A,V,f)$,集合 X 的 R 下近似集为 $\underline{R}(X)$,X 的 R 上近似集为 $\overline{R}(X)$,对论域 U 进行划分形成的等价类为 $U/R=\{E_1,E_2,\cdots,E_n\}$,令属性增加之后,集合 X 的 R 上近似集、下近似集分别为 $\overline{R^{\vee}}(X)$ 和 $\underline{R^{\vee}}(X)$,那么 $\underline{R^{\vee}}(X)\supseteq\underline{R}(X)$,$\overline{R^{\vee}}(X)\subseteq\overline{R}(X)$。

证明:令 $\forall x\in\underline{R}(X)$,设由 x 形成的等价类为 $E_i((i=1,2,\cdots,n)$,那么 $E_i\subseteq X$,当属性增加时,等价类 E_i 被细划成的新等价类为 $E_{ik}(k=1,2,\cdots j_i)$,因此 $E_{ik}\subseteq E_i$,那么 $E_{ik}\subseteq X$,所以 $\forall x\in\underline{R^{\vee}}(X)$。则 $\underline{R^{\vee}}(X)\supseteq\underline{R}(X)$。令 $\forall x\in\overline{R^{\vee}}(X)$,此时,集合 X 和由 x 形成的等价类 E_{ik} 的交集不为空,即 $E_{ik}\cap X\neq\varnothing$,由于属性增加时 $E_{ik}\subseteq E_i$,所以 $E_i\cap X\neq\varnothing$,所以 $\forall x\in\overline{R}(X)$,即 $\overline{R^{\vee}}(X)\subseteq\overline{R}(X)$。综上可得,$\underline{R^{\vee}}(X)\supseteq\underline{R}(X)$,$\overline{R^{\vee}}(X)\subseteq\overline{R}(X)$。

定理 2 给定信息系统 $S=(U,A,V,f)$,等价类 $E_i(i=1,2,\cdots,n)$ 是对论域 U 进行划分的等价类。属性增加后,等价类 E_i 被划分成新的等价类 $E_{ik}(k=1,2,\cdots j_i)$。若 $E_i\subseteq\overline{R}(X)-\underline{R}(X)$,(1)如果 $E_{ik}\cap X=\varnothing$,则 $\overline{R^{\vee}}(X)=\overline{R}(X)-E_{ik}$;(2)如果 $E_{ik}\subseteq X$,则 $\underline{R^{\vee}}(X)=\underline{R}(X)\cup E_{ik}$;(3)如果 $E_{ik}\subseteq\overline{R}(X)-\underline{R}(X)$,则 $\underline{R^{\vee}}(X)=\underline{R}(X)$,$\overline{R^{\vee}}(X)=\overline{R}(X)$。

定理的证明过程与定理 1 的证明过程类似,在此不再论述,下同。

1.2 算法思想及其描述

1.2.1 属性增加时,动态获取近似集的算法思想

首先判断每个等价类中对象对应的增加属性的属性值是否相等,若每个等价类中对象的增加属性的属性值全部都相等,那么等价类不进行细化,此时,近似集不发生变化。若每个等价类中对象的增加属性的属性值有不相等的情况,那么,此时原等价类会被划分得到新等价类。然后根据原等价类与集合 X 之间的关系,分别分析判断新等价类与 X 的关系,再根据等价类与原近似集得到动态变化之后的近似集。若原等价类与集合 X 关系是包含关系或者相交为空集,则不需要考虑原等价类对近似集的影响。

1.2.2 算法的流程

属性增加时动态获取近似集流程图如图 1 所示。

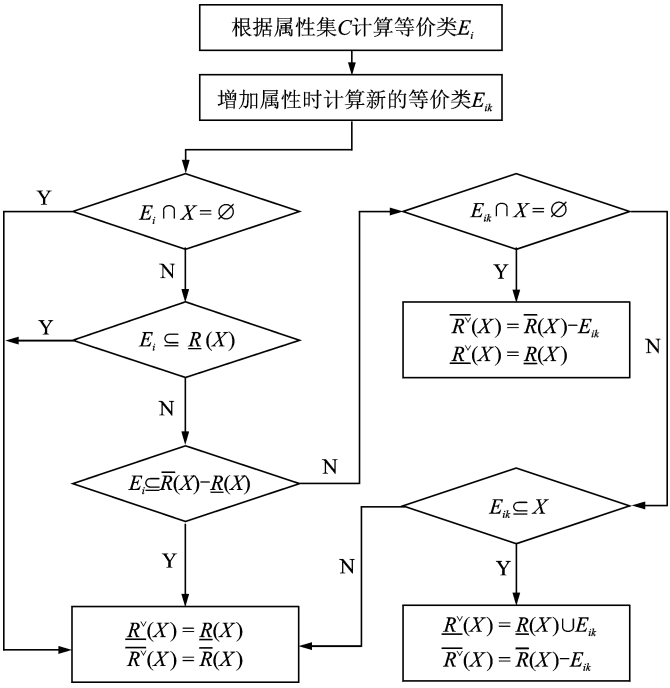

图 1 属性增加时动态获取近似集流程图

Fig. 1 Process of dynamical updating approximations while adding attribute

1.2.3 算法的时间复杂度分析

对于信息系统 S 来说,$|U|$ 表示信息系统中对象的个数,$|C|$ 表示信息系统中属性的个数,信息系统中等价类的个数为 $2^{|C|}$ 个。集合 X 中对象的个数表示为 $|X|$。增加属性时采用传统的方法计算等价类的最坏时间复杂度为 $O(|U||C|\log|U|)$,计算近似集的最坏时间复杂度为 $O(2^{|C|}|X|)$,所以总的时间复杂度为 $O(|U||C|\log|U|+2^{|C|}|X|)$。采用动态获取算法,令 $t=2^{|C|}$,当属性增加后,$|U|$ 个对象被拆分成 t 个等价类,分别为 $E_1,E_2,\cdots,E_t$,且 $E_1 \cup E_2 \cup \cdots \cup E_t = U$,每个等价类中对象的个数分别为 $n_1,n_2,\cdots,n_t$,计算等价类时间复杂度为 $O(n_1\log n_1)+O(n_2\log n_2)+\cdots+O(n_t\log n_t)$,计算近似集的时间复杂度为 $O(2^{|C|}|X|)$,此时,动态获取近似集总时间复杂度为 $O(n_1\log n_1+n_2\log n_2+\cdots+n_t\log n_t+2^{|C|}|X|)$。对两个算法时间复杂度比较,$n_1\log n_1+n_2\log n_2+\cdots+n_t\log n_t=\log n_1^{n_1}+\log n_2^{n_2}+\cdots+\log n_t^{n_t}<\log|U|^{n_1+n_2+\cdots+n_t}=\log|U|^{|U|}=|U|\log|U|<|U||C|\log|U|$。因此,动态获取算法的总的时间复杂度比原始算法总的时间复杂度要低。

2 属性减少时近似集动态获取方法

属性减少时,等价关系对论域的划分可能发生变化,上近似集和下近似集可能会随着属性的减少发生变化。此时,原信息系统中不同的等价类可能合并得到新的等价类,通过判断原等价类与近似集之间的关系,得到动态更新之后新的上近似集和下近似集。这样可以减少重复计算所花费的时间,提高效率。

2.1 减少单个属性时近似集获取方法

当属性减少时,原等价类有被合并的可能,因此上近似集有增加的可能,下近似集有减少的可能。下面给出经典粗糙集模型中,当属性减少时近似集动态获取相关定理。

定义 5 设信息系统 $S=(U,A,V,f)$,对论域 U 进行划分形成的等价类为 $U/R=\{E_1,E_2,\cdots,E_n\}$,属性减少前集合 X 的 R 上近似集和 R 下近似集分别记为$\overline{R}(X)$和$\underline{R}(X)$。当属性减少之后,如果等价类 $E_i(i=1,2,\cdots,n)$和 $E_j(j=1,2,\cdots,n)$被粗化,那么,令等价类 E_i 和 E_j 被粗化之后形成的等价类为 $E_k^{\wedge}(k=1,2,\cdots,l)$,属性减少后集合 X 的 R 上近似集和 R 下近似集分别记为$\overline{R^{\wedge}}(X)$和$\underline{R^{\wedge}}(X)$。

推论 2 设信息系统 $S=(U,A,V,f)$,对论域 U 进行划分形成的等价类为 $U/R=\{E_1,E_2,\cdots,E_n\}$,等价类 $E_i(i=1,2,\cdots,n)$和 $E_j(j=1,2,\cdots,n)$为任意的两个等价类,$f:U\times A\to V$ 表示为一个信息函数,C 为信息系统中条件属性,d 为减少的单个属性。(1)如果 $f_{\forall x_i\in E_i}(d)=f_{\forall x_j\in E_j}(d)(i,j=1,2,\cdots,n,i\neq j)$,则等价类 E_i 和 E_j 不会发生粗化。(2)如果 $f_{\forall x_i\in E_i}(C-\{d\})=f_{\forall x_j\in E_j}(C-\{d\})\ (i,j=1,2,\cdots,n,i\neq j)$,此时,等价类 E_i 和 E_j 被粗化,令 E_i 和 E_j 被粗化之后形成的等价类为 $E_k^{\wedge}(k=1,2,\cdots,l)$,则 $E_k^{\wedge}=E_i\cup E_j$。

定理 3 在经典粗糙集模型中,给定信息系统 $S=(U,A,V,f)$,集合 X 的 R 上近似和 R 下近似集 $\underline{R}(X)$和$\overline{R}(X)$,信息系统中论域 U 关于条件属性所形成的等价类为 $U/R=\{E_1,E_2,\cdots E_n\}$,属性减少之后集合 X 的 R 上近似集和 R 下近似集分别记为$\overline{R^{\wedge}}(X)$和$\underline{R^{\wedge}}(X)$。那么,$\underline{R^{\wedge}}(X)\subseteq\underline{R}(X)$,$\overline{R^{\wedge}}(X)\supseteq\overline{R}(X)$。定理的证明过程与属性增加时定理 1 的证明过程类似,在此,不再论述,以下定理证明过类似。

定理 4 设信息系统 $S=(U,A,V,f)$,对论域 U 进行划分形成的等价类为 $U/R=\{E_1,E_2,\cdots,E_n\}$,当属性减少后,如果原信息系统中等价类 $E_i(i=1,2,\cdots,n)$和 $E_j(j=1,2,\cdots,n)$被粗化,那么,令等价类 E_i 和 E_j 被粗化之后形成的等价类为 $E_k^{\wedge}(k=1,2,\cdots,l)$,动态变化后 X 的负域为 $\mathrm{negr}^{\wedge}(X)$。属性减少之前,当 $E_i\cap X=\varnothing$时(1)如果 $E_j\subseteq\underline{R}(X)$,那么$\overline{R^{\wedge}}(X)=\overline{R}(X)\cup E_i$,$\underline{R^{\wedge}}(X)=\underline{R}(X)-E_j$;(2)如果 $E_j\subseteq\overline{R}(X)-\underline{R}(X)$,那么$\overline{R^{\wedge}}(X)=\overline{R}(X)\cup E_i$,$\underline{R^{\wedge}}(X)=\underline{R}(X)$;(3)如果 $E_j\cap X=\varnothing$,那么$\underline{R^{\wedge}}(X)=\underline{R}(X)$,$\overline{R^{\wedge}}(X)=\overline{R}(X)$。

定理 5 **设信息系统** $S=(U,A,V,f)$,对论域 U 进行划分形成的等价类为 $U/R=\{E_1,E_2,\cdots,E_n\}$,当属性减少后,如果原信息系统中 $E_i(i=1,2,\cdots,n)$和 $E_j(j=1,2,\cdots,n)$被粗化,那么,令等价类 E_i 和 E_j 被粗化之后形成的等价类为 $E_k^{\wedge}(k=1,2,\cdots,l)$,动态变化后 X 的边界域为 $\mathrm{bnr}^{\wedge}(X)$。属性减少之前,当 $E_i\subseteq\overline{R}(X)$时(1)如果 $E_j\subseteq\mathrm{bnr}(X)$,那么$\overline{R^{\wedge}}(X)=\overline{R}(X)$;$\underline{R^{\wedge}}(X)=\underline{R}(X)-E_i$;(2)如果 $E_j\subseteq\underline{R}(X)$,那么$\overline{R^{\wedge}}(X)=\overline{R}(X)$;$\underline{R^{\wedge}}(X)=\underline{R}(X)$。

2.2 算法思想及其描述

2.2.1 属性减少时,动态获取近似集算法思想

已知原信息系统的上近似集、下近似集和等价类,属性减少时,可以根据判断等价类 E_i 中所有元素对应动态变化后属性的属性值与等价类 E_j 中所有元素对应动态变化后属性的属性值是否相等,得知原等价类 E_i 和 E_j 之间是否有等价关系。如果存在等价关系,那么新等价类 $E_k^{\wedge}$ 可以表示成等价类 E_i 和 E_j 的并,如果有两个以上的等价类存在等价关系,那么新等价类 $E_k^{\wedge}$ 则是所有存在等价关系的等价类的并,否则等价类不发生变化。当属性减少后有新的等价类 $E_k^{\wedge}$ 产生时,根据相关定理动态获取近似集。

2.2.2 算法的流程

属性减少时动态获取近似集流程图如图 2 所示。

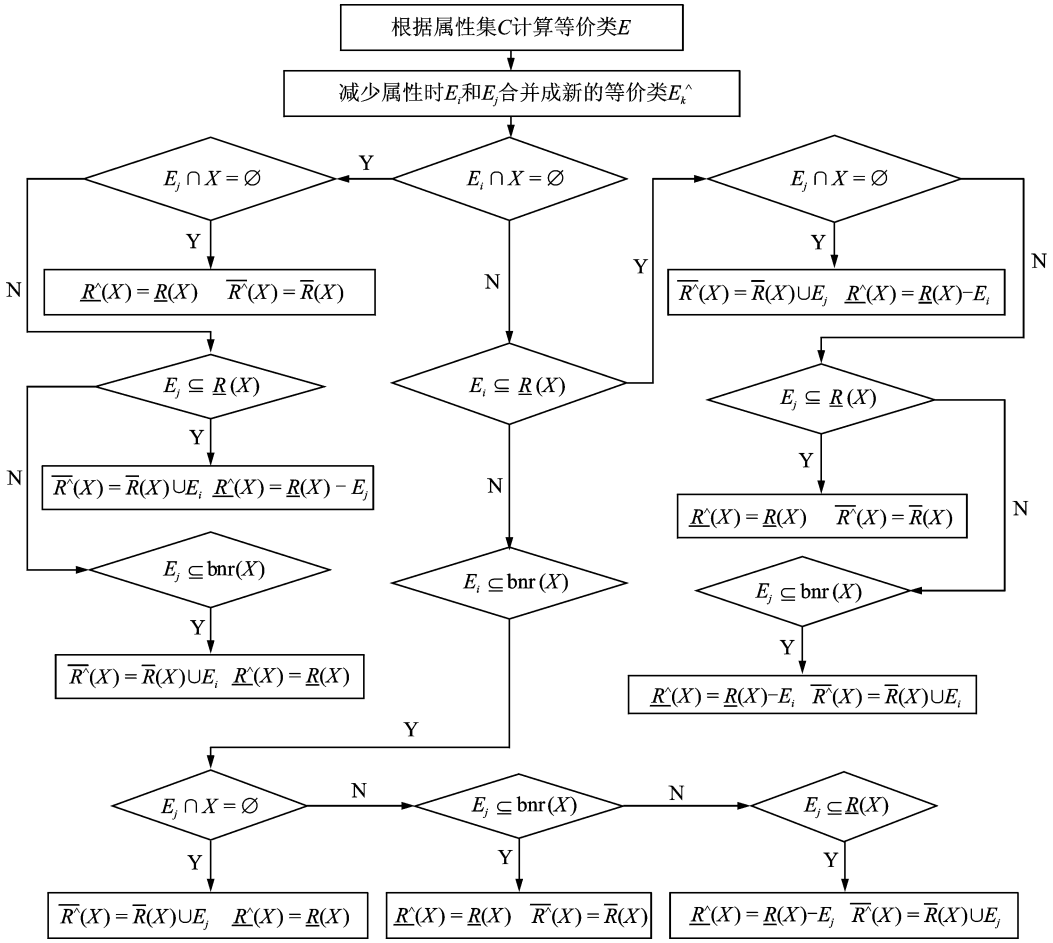

图 2 属性减少时动态获取近似集流程图

Fig. 2 Process of dynamical updating approximations while deleting attribute

2.2.3 算法的时间复杂度分析

设 S 表示信息系统,$|U|$ 表示对象个数,$|C|$ 表示属性的个数,则信息系统中等价类的个数为 $2^{|C|}$($|U| \geqslant 2^{|C|}$)。集合 X 中对象的个数表示为 $|X|$。当减少属性时,形成的等价类个数为 t($t \leqslant 2^{|C|}$)个,采用原始算法,计算等价类的最坏的时间复杂度为 $O(|U||C|\log|U|)$,计算近似集的时间复杂度为 $O(t|X|)$,故更新近似集总的时间复杂度为 $O(|U||C|\log|U|+t|X|)$;采用动态获取算法而言,属性减少后,$2^{|C|}$ 个等价类合并成 t 个等价类,计算等价类时间复杂度为 $O(|C|^2 2^{|C|})$,计算近似集的时间复杂度为 $O(t|X|)$,此时,采用动态获取算法总的时间复杂度为 $O(|C|^2 2^{|C|}+t|X|)$。因为,$|C||U|\log|U|+t|X| \geqslant |C|2^{|C|}\log 2^{|C|}+t|X| = |C|^2 2^{|C|}+t|X|$。因此动态更新算法的时间复杂度比原始算法低。

3 仿真实验及结果分析

算法的仿真实验环境为:Windows XP 操作系统,Intel(R) Core(TM)2 Duo CPU 处理器,2 GB 内存,在 Visual Studio 2008 开发平台中实现仿真实验。为了验证属性增加时所提出算法的有效性,分别

用传统算法和所提出的算法从 UCI 数据库中选取 10 个数据集进行仿真实验。实验过程中,随机选取数据集中对象个数的 10%作为集合 X,从每个数据集中选取 4 个条件属性作为原始属性集。通过对信息系统中单个属性增加进行仿真实验,对每个数据集进行 10 次动态获取计算,得到在每个数据集上获取近似集的时间,然后求出动态更新近似集的平均时间,得到如表 1 所示的属性增加时原始算法和动态更新算法计算近似集消耗时间比较表。

表 1 属性增加时两种算法计算近似集时间比较表

Table 1 Comparison of time for computing approximations between two algorithms while adding attribute

编号	数据集	对象数	属性数	原始算法耗时/s	动态更新算法耗时/s
1	lenses	24	5	0.000 251	0.000 152
2	zoo	101	17	0.000 644	0.000 355
3	Monk's problems	345	7	0.002 942	0.001 634
4	tic-tac-toe	958	9	0.006 403	0.003 694
5	contraceptive-method-choice	1 473	10	0.014 066	0.008 257
6	car evaluation	1 728	7	0.013 899	0.008 03
7	chess(king-rook-vs-king-pawn)	3 196	37	0.027 419	0.020 315
8	statlog	4 435	37	0.065 024	0.027 997
9	nursery	12 960	9	0.143 745	0.040 126
10	poker-hand-training-true	25 010	11	0.240 982	0.118 988

属性增加时,原始算法与动态获取算法计算近似集消耗时间的比较,如图 3 所示。由表 1 和图 3 可知,当数据集比较小时,两种算法的时间效率相差不大,花费的时间都比较少。当数据集变大时,动态获取算法的耗时增长缓慢,而原始算法的耗时增长速度较快。因此,动态获取近似集算法的时间效率要优于原始算法获取近似集的时间。分析原因可知,在信息系统动态变化下,当属性动态增加时,采用原始的算法求近似集需要重新计算求解出所有的等价类,在计算的过程中花费了较多的时间,对于动态获取算法而言,只需要对每个对象增加属性的属性值进行比较,不需要对原属性集的属性值进行比较,提高了效率。

为了验证属性减少时近似集动态获取算法的有效性,以下实验分别用动态获取算法和原始算法在 10 个数据集上进行计算近似集。实验过程中,选取每个数据集的 5 个条件属性作为原始属性集,选取原始属性集中的其中一个属性作为减少的属性,然后动态获取近似集。属性减少时动态获取算法和原始算法的时间比较如表 2 所示。属性减少时原始算法与动态获取算法计算近似集所花费时间的比较,如图 4 所示。由图 4 和表 2 可知,随着属性的减少,当数据集较小时,原始算法和动态获取算法计算近似集的时间相差不多。当数据集较大时,动态获取算法花费时间增长缓慢,而原始算法增长速度较快,因此,动态获取算法的时间效率要优于原始算法。在信息系统属性动态减少时,原始算法仍然需要采用传统方法进行计算等价类,整个计算过程花费了大量的时间,对于动态获取算法而言,只需对等价类中的一个对象进行比较,判断等价类之间有新的等价类产生。如果等价类的个数较少时,等价类之间进行比较得到新等价类的次数就比较少,更新近似集的效率就会有所提高。由图 4 可知,当数据集较大时,动态获取算法计算近似集效率较高。

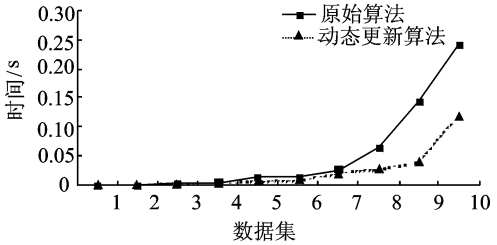

图 3 属性增加时原始算法与动态获取算法计算近似集时间比较图

Fig. 3 Comparison of time for computing approximations between original algorithm and dynamical updating algorithm while adding attribute

表 2 属性减少时两种算法计算近似集时间比较表

Table 2 Comparison of time for computing approximations between two algorithms while deleting attribute

编号	数据集	对象数	属性数	原始算法耗时/s	动态更新算法耗时/s
1	lenses	24	5	0.000 207	0.000 145
2	zoo	101	17	0.000 657	0.000 285
3	Monk's problems	345	7	0.002 615	0.001 294
4	tic-tac-toe	958	9	0.006 635	0.003 025
5	contraceptive-method-choice	1 473	10	0.010 160	0.006 737
6	car evaluation	1 728	7	0.012 619	0.003 982
7	chess(king-rook-vs-king-pawn)	3 196	37	0.028 526	0.011 233
8	statlog	4 435	37	0.061 767	0.035 957
9	nursery	12 960	9	0.133 672	0.028 378
10	poker-hand-training-true	25 010	11	0.252 509	0.070 815

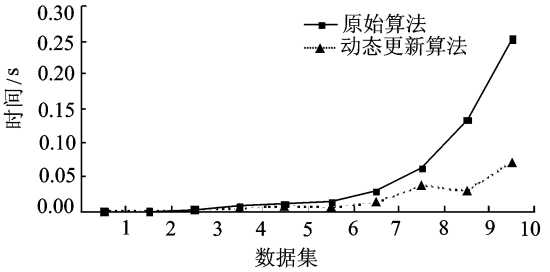

图 4 原始算法与动态获取算法在属性减少时计算近似集所花费的时间比较

Fig. 4 Comparison of time for computing approximations between original algorithm and dynamical updating algorithm while deleting attribute

4 结束语

信息系统是在不断变化的,采用传统方法划分等价类,会花费大量时间进行重复计算,效率较低。通过对原信息系统中已知的信息分析,可不需要对等价类进行重新计算,并根据对等价类的判断,得到经典粗糙集模型中属性增加或者减少时近似集变化的相关定理,提出了属性增加和减少时经典粗糙集模型中动态获取近似集的算法。最后通过 UCI 数据集测试了算法的有效性,并且在效率上有所提高,但是本文并没有考虑信息系统含有决策属性以及多属性变化情况下,近似集和规则的动态获取,这些是今后研究的重点。

参考文献:

[1] Pawlak Z. Rough sets[J]. International Journal of Computer and Information Sciences, 1982, 11: 341-356.

[2] 王国胤. Rough 集理论与 Rough 集知识获取[M]. 西安:西安交通大学出版社,2001:11-15.
Wang Guoyin. Rough set theory and rough set knowledge acquisition[M]. Xi′an : Xi′an Jiaotong University Press,2001:11-15.

[3] 张文修,梁怡,吴伟志,等. 信息系统与知识发现[M]. 北京:科学出版社,2003:23-29.
Zhang Wenxiu, Liang Yi, Wu Weizhi. Information system and knowledge discovery[M]. Beijing :Science Press,2003.

[4] 李元诚,方廷健. 基于粗糙集理论的支撑向量机预测方法研究[J]. 数据采集与处理,2003,18(2):199-203.
Li Yuancheng, Fang Tingjian. Study of forcasting algorithm for support vector machines based on rough sets[J]. Journal of Data Acquisition and Processing,2003,18(2):199-203.

[5] 崔建国,宋博翰,董世良,等. 基于邻域粗糙集的航空发电机健康诊断方法[J]. 数据采集与处理,2012,27(1):80-84.
Cui Jianguo, Song Bohan, Dong Shiliang. Health diagnosis of aero-generator based on neighborhood rough sets theory[J]. Journal of Data Acquisition and Processing, 2012, 27(1):80-84.

[6] 徐立中,王慧敏,刘美林,等. 粗糙集理论在图像增强中的应用[J]. 数据采集与处理,1999,14(3):307-310.
Xu Lizhong, Wang Huimin, Liu Meilin, et al. An image enhancing method based on rough sets[J]. Journal of Data Acquisition and Processing, 1999, 14(3):307-310.

[7] Chan C C. A rough set approach to attribute generalization in data mining[J]. Information Sciences, 1998, 107: 177-194.

[8] Li Tianrui, Ruan Da, Greet W, et al. A rough sets based characteristic relation approach for dynamic attribute generalization in data mining[J]. Knowledge-based Systems,2007,20(5):485-494.

[9] 王磊,李天瑞,刘清,等. 对象集变化时近似集动态维护的矩阵方法[J]. 计算机研究与发展,2013,50(9):1992-2004,
Wang Lei, Li Tianrui, Liu Qing, et al. A matrix-based approach for maintenance of approximations under the variation of object set[J]. Journal of Computer Research and Development,2013,50(9):1992-2004.

[10] Li Tianrui, Xu Yang. A generalized rough set approach to attribute generalization in data mining[J]. Journal of Southwest Jiaotong University, 2000, 8(1): 69-75.

[11] Li Tianrui, Ma Jun, Xu Yang, et al. An approach to attribute generalization in incomplete information system[C]// Proceedings of International Conference on Machine Learning and Cybernetics. Xi′an, China: [s. n.],2003:1678-1691.

[12] 曾安平,李天瑞,罗川. 高斯核模糊粗糙集中对象集变化时近似集增量更新方法研究[J]. 计算机科学,2013,40(7):172-177.
Zeng Anping, Li Tianrui, Luo Chuan. Incremental approach for updating approximations of gaussian kernelized fuzzy rough sets under variation of object set[J]. Journal of Computer Science,2013,40(7):172-177.

[13] Chen Hongmei, Li Tianrui, Ruan Da, et al. A rough-set-based incremental approach for updating approximations under dynamic maintenance enviroments[J]. IEEE Transaction on Knowledge and Data Engineering ,2013,25(2):274-284

[14] Cheng Yi. The incremental method for fast computing the rough fuzzy approximation[J]. Data and Knowledge Engineering, 2011,70(1):84-100.

[15] Zhang Junbo, Li Tianrui, Ruan Da, et al. Rough sets based matrix approaches with dynamic attribute variation in set-valued information systems[J]. International Journal of Approximation Reasoning,2012,153(4):620-635.

[16] Zhang Junbo, Li Tianrui, Ruan Da, et al. A parallel method for computing rough set approximations[J]. Information Sciences,2012,194(1):209-223.

[17] Liu Dun, Li Tianrui, Zhang Junbo, A rough set-based incremental approach for learning knowledge in dynamic incomplete information systems[J]. International Journal of Approximate Reasoning, 2014,55(8): 1764-1786.

作者简介:

胡成祥(1984-),男,助教,研究方向:智能信息处理、粗糙集理论,E-mail:chengxiang0550@163.com。

赵瑞斌(1983-),男,讲师,研究方向:计算图形学、并行计算、大数据处理,E-mail:zhao_rui_bin@163.com。