缺失数据数据集的组增量式特征选择

王 锋 魏 巍
(山西大学计算机与信息技术学院 太原 030006)

摘 要 实际应用中获取到的数据集通常是动态增加的,且随着数据获取工具的迅速发展,新数据通常会一组一组地增加。为此,针对含有缺失数据的动态数据集,基于粗糙集理论,提出了一种组增量式的粗糙特征选择算法。首先分析、证明了信息熵的组增量计算公式,并以信息熵作为特征重要度的度量,在此基础上设计了基于信息熵的可有效处理含有缺失数据的动态数据集的组增量式特征选择算法。实验结果进一步证明了新算法的可行性和高效性。

关键词 动态数据集,缺失数据,信息熵,组增量特征选择
中图法分类号 TP181 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2015.7.061

Group Feature Selection Algorithm for Data Sets with Missing Data

WANG Feng WEI Wei
(School of Computer and Information Technology, Shanxi University, Taiyuan 030006, China)

Abstract Many real data increase dynamically in size. With the rapid development of data processing tools, new data are usually increased in groups. In this paper, based on rough set theory, a group rough feature selection algorithm was proposed to deal with dynamic data sets with missing data. Firstly, the group incremental mechanism of information entropy was analyzed, and then significance of feature was defined based on the mechanism. On this basis, a group feature selection algorithm was constructed, which can be used to deal with dynamic data sets with missing data effectively. Experimental results show that the new algorithm is feasible and efficient.

Keywords Dynamic data sets, Missing data, Information entropy, Group feature selection

1 引言

信息技术的迅速发展和数据获取工具的进步,使得各种各样的观测设备、实验仪器以及网络传感器等都以爆炸式增长的趋势更新着对客观世界的描述,从而在许多实际应用领域中获取到的数据都是动态更新的。尤其是目前大数据时代的到来更进一步加快了数据更新的速度,甚至于无法对其估计。新数据的大量产生带来了数据中信息的动态更新,这也使得我们应及时而准确地去获取最新的有用知识。面临规模庞大且快速更新的动态数据集,传统的处理静态数据集的数据挖掘方法显然不能满足处理快速更新的信息所需的实效性。因此,针对动态更新的数据集,运用动态的数据挖掘方法及时地分析并获取有用信息就显得尤为重要,探索和研究高效的动态数据挖掘算法也迅速成为数据挖掘领域中的研究热点之一。

对数据挖掘技术更为深入的研究表明,数据预处理对更有效地使用数据挖掘工具并成功进行知识发现是至关重要的。特征选择是一个目前已被广泛采用的数据预处理技巧[1-4],现已被广泛用于文本分类、图像检索、入侵检测以及基因分析等许多应用领域[5-7]。针对动态更新的数据集,一些研究者对增量式的特征选择已经进行了相应的探索[8-10]。Liu使用布尔矢量来表示特征子集,并形成一个布尔矩阵。当新数据增加后,通过更新原有的布尔矩阵来求解新的特征选择结果[11]。但是该算法只处理无类信息的数据表,属于一类无监督的增量式特征选择算法。针对带有类信息的数据表,Orlowska和Yang等分别基于粗糙集理论中区分矩阵的概念,通过更新和化简区分矩阵,给出了一类有监督的增量式特征选择算法[12-14]。为更高效地更新特征子集,Hu等提出了基于正区域的增量式特征选择算法[15],Liang等提出了基于信息熵的增量式特征选择算法[16,17]。然而,对给定的动态数据表,上述算法都属于单增量的更新方式,即每次处理单个新增的数据对象。数据获取设备的迅速发展,加之大数据时代的到来,使得在更多的实际应用中新数据会一组一组增加,Liang等通过分析3种常用信息熵的组增量更新机制,提出了一种组增量的粗糙特征选择算法,其可高效地一次处理一组新增的数据[18]。

数据取值缺失的现象在许多实际应用中是广泛存在的,为此探索面向含有缺失数据数据集的特征选择也已成为数据挖掘领域的一个热点问题[19-21]。一些研究者已经展开了深入的研究,并取得了可观的研究成果[22-24]。随着数据更新速度的不断加快,以及增量式处理技巧的引入和使用,如何高效地求解含有缺失数据的动态数据集的特征子集便成为特征选

到稿日期:2014-07-29 返修日期:2014-10-22 本文受国家自然科学基金项目(61402272)资助。
王 锋(1984-),女,博士,讲师,主要研究方向为粗糙集、特征选择,E-mail:sxuwangfeng@126.com;魏 巍(1980-),男,博士,副教授,主要研究方向为粗糙集、粒度计算。

择研究中的又一个热点研究问题[25]。本文基于粗糙集理论，使用信息熵来度量候选特征的重要度，设计了一种面向缺失数据数据集的组增量特征选择算法。由于信息熵的计算对求解候选特征的重要度显然是非常重要的，因此文中首先分析并证明了含有缺失数据数据集上信息熵的组增量式更新机制。在此基础上，重新定义特征重要度，进而设计了基于信息熵的组增量特征选择算法。为验证新算法的可行性和高效性，文中选取了4组UCI数据库中含有缺失数据取值的数据集，并分别与基于信息熵的非增量算法和单增量算法进行了实验比较和测试，实验结果进一步证明了新算法的有效性。

本文第2节基于粗糙集理论介绍了含有缺失数据数据集的符号表示和互补信息熵的定义；第3节分别分析并证明了互补信息熵的单增量更新机制和组增量更新机制；第4节重新定义了特征重要度的度量，并分别给出了面向缺失数据数据集的单增量式特征选择算法和组增量式特征选择算法；第5节选取了4组UCI数据集进行了实验测试和分析；最后总结全文。

2 基本概念

基于粗糙集模型，本节介绍含有缺失数据数据集的符号表示[23,26]。

粗糙集理论中，给定数据表 $S=(U,A)$ 被称为一个信息系统。其中 U 表示数据表中所有数据对象的集合，称为论域；A 表示数据表中所有特征的集合，称为属性。对任意的 $a\in A$，有 $a:U\to V_a$，其中 V_a 称为属性(特征)a 的值域；对任意 $a\in A, u\in U$，有 $f(u,a)\in V_a$，其中 $f(u,a)$ 是一个信息函数，它对一个数据对象的每一个属性赋予一个信息值；如果至少有一个属性 $a\in A$ 使得 V_a 中含有空值，即数据表 S 中含有缺失的数据取值，则 S 称为一个非完备信息系统，并使用 $*$ 表示数据表中的空值(缺失的数据取值)。

设 $P\subseteq A$，由 P 诱导的相容关系定义如下：$SIM(P)=\{(u,v)\in U\times U \mid \forall a\in P, f(u,a)=f(u,a)$ 或 $f(u,a)=*$ 或 $f(v,a)=*\}$。

$S_P(u)=\{v\in U|(u,v)\in SIM(P)\}$，$S_P(u)$ 表示与 u 可能不可区分的数据对象的最大集合。$U/SIM(P)$ 表示数据表上的一个分类，其中的元素称为相容类。

一个带有类标签的数据表表示为 $S=(U,C\cup D)$，称为决策表，其中 D 表示类标签的一列，称为决策属性；C 表示所有特征的集合，称为条件属性集。含有缺失数据取值的决策表称为非完备决策表。

定义1 令 $S=(U,C\cup D)$ 是一个非完备决策表，$P\subseteq C$，则互补信息熵的条件熵定义为

$$E(D|P)=\frac{1}{|U|^2}\sum_{i=1}^{|U|}(|S_P(u_i)|-|S_P(u_i)\cap S_D(u_i)|)$$

3 信息熵的增量式更新机制

对于给定的非完备决策表，基于定义1中的条件信息熵，本节主要介绍信息熵的增量式更新机制。通过分析新数据增加到给定数据集中后相容类的变化，分别给出了信息熵的单增量式更新机制和组增量式更新机制。具体介绍如下。

3.1 互补信息熵的单增量式更新机制

定理1 令 $S=(U,C\cup D)$ 是一个非完备决策表，$P\subseteq C$，且有 $U/SIM(P)=\{S_P(u_1),S_P(u_2),\cdots,S_P(u_{|U|})\}$ 和 $U/SIM(D)=\{S_D(u_1),S_D(u_2),\cdots,S_D(u_{|U|})\}$。论域 U 上 D 关于 P 的互补信息熵记为 $E_U(D|P)$。当有新的数据对象 x 增加到数据表 S 中后，假设在条件属性集 C 上与 x 满足相容关系的数据对象集为 X，而在决策属性 D 上与 x 满足相容关系的数据对象集为 Y，那么，论域 $U\cup\{x\}$ 上的互补信息熵为

$$E_{U\cup\{x\}}(D|P)=\frac{1}{(|U|+1)^2}(|U|^2E_U(D|P)+2|X-Y|)$$

证明：当增加新数据 x 后，x 与原来论域 U 中数据对象间的相容关系可具体分为以下4种情况：

(1)对 $\forall y\in U-X\cup Y$，x 与 y 在 C 和 D 上均不满足相容关系；

(2)对 $\forall y\in Y-X$，x 与 y 在 C 上不满足相容关系，而在 D 上满足相容关系；

(3)对 $\forall y\in X-Y$，x 与 y 在 C 上满足相容关系，而在 D 上不满足相容关系；

(4)对 $\forall y\in X\cap Y$，x 与 y 在 C 和 D 上都满足相容关系。

令 $m=|U-X\cup Y|$，$U\cup\{x\}/SIM(P)=\{S_P'(u_1),S_P'(u_2),\cdots,S_P'(u_{|U|}),S_P'(x)\}$，$U\cup\{x\}/SIM(D)=\{S_D'(u_1),S_D'(u_2),\cdots,S_D'(u_{|U|}),S_D'(x)\}$。又由于 $(U-X\cup Y)\cup(Y-X)\cup(X-Y)\cup(X\cap Y)=U$，基于上述4种情况的讨论，论域 $U\cup\{x\}$ 上的互补信息熵分析如下：

$$\begin{aligned}
&E_{U\cup\{x\}}(D|P)\\
&=\frac{1}{(|U|+1)^2}\Big(\sum_{i=1}^{m}(|S_P'(u_i)|-|S_P'(u_i)\cap S_D'(u_i)|)+\sum_{i=m}^{|U-X|}(|S_P'(u_i)|-|S_P'(u_i)\cap S_D'(u_i)|)+\sum_{i=|U-X|}^{m+|Y|}(|S_P'(u_i)|-|S_P'(u_i)\cap S_D'(u_i)|)+\sum_{i=m+|Y|}^{|U|}(|S_P'(u_i)|-|S_P'(u_i)\cap S_D'(u_i)|)+|S_P'(x)|-|S_P'(x)\cap S_D'(x)|\Big)\\
&=\frac{1}{(|U|+1)^2}\Big(\sum_{i=1}^{m}(|S_P(u_i)|-|S_P(u_i)\cap S_D(u_i)|)+\sum_{i=m}^{|U-X|}(|S_P(u_i)|-|S_P(u_i)\cap(S_D(u_i)\cup\{x\})|)+\sum_{i=|U-X|}^{m+|Y|}(|S_P(u_i)\cup\{x\}|-|(S_P(u_i)\cup\{x\})\cap(S_D(u_i)\cup\{x\})|)+\sum_{i=m+|Y|}^{|U|}(|S_P(u_i)\cup\{x\}|-|(S_P(u_i)\cup\{x\})\cap S_D(u_i)|)+|X-X\cap Y|\Big)\\
&=\frac{1}{(|U|+1)^2}\Big(\sum_{i=1}^{m}(|S_P(u_i)|-|S_P(u_i)\cap S_D(u_i)|)+\sum_{i=m}^{|U-X|}(|S_P(u_i)|-|S_P(u_i)\cap S_D(u_i)|)+\sum_{i=|U-X|}^{m+|Y|}(|S_P(u_i)|-|S_P(u_i)\cap S_D(u_i)|)+\sum_{i=m+|Y|}^{|U|}(|S_P(u_i)|-|S_P(u_i)\cap S_D(u_i)|+1)+|X-Y|\Big)\\
&=\frac{1}{(|U|+1)^2}\Big(\sum_{i=1}^{|U|}(|S_P(u_i)|-|S_P(u_i)\cap S_D(u_i)|)+2|X-Y|\Big)=\frac{1}{(|U|+1)^2}(|U|^2E_U(D|P)+2|X-Y|)
\end{aligned}$$

由定理1可得，当新数据增加到给定的含有缺失数据的数据集中后，通过分析与新数据 x 满足相容关系的数据集，即可计算得到新数据集上的互补信息熵，从而避免了在新数据集上重新求解熵的重复计算。

3.2 互补信息熵的组增量式更新机制

当有一组新数据在同一时刻增加到给定数据集中后，使

用上述的单增量计算公式显然会比较耗时。为进一步提高计算效率，本节分析互补信息熵的组增量计算公式。

定理 2 令 $S=(U,C\cup D)$是一个非完备决策表，$P\subseteq C$，且有 $U/SIM(P)=\{S_P(u_1),S_P(u_2),\cdots,S_P(u_{|U|})\}$和 $U/SIM(D)=\{S_D(u_1),S_D(u_2),\cdots,S_D(u_{|U|})\}$。论域 U 上 D 关于 P 的互补信息熵记为 $E_U(D|P)$。当有一组新的数据对象 V 增加到数据表 S 中后，有 $V/SIM(P)=\{S_P(v_1),S_P(v_2),\cdots,S_P(v_{|V|})\}$和 $V/SIM(D)=\{S_D(v_1),S_D(v_2),\cdots,S_D(v_{|V|})\}$。假设在新论域 $U\cup V$ 上，$X_1\subseteq U$ 和 $X_2\subseteq V$ 分别表示在 C 上相容类发生变化的数据对象集，$Y_1\subseteq U$ 和 $Y_2\subseteq V$ 分别表示在 D 上相容类发生变化的数据对象集，那么新论域 $U\cup V$ 上的互补信息熵为

$$E_{U\cup V}(D|P)=\frac{1}{|U\cup V|^2}(|U|^2E_U(D|P)+|V|^2E_V(D|P)+\Delta)$$

其中，$\Delta=\sum_{i=m}^{m+|X_1|}|X_P(u_i)|+\sum_{i=m'}^{m'+|X_2|}|X_P(v_i)|-\sum_{i=|U-Y_1|}^{m+|X_1|}|X_P(u_i)\cap X_D(u_i)|-\sum_{i=|V-Y_2|}^{m'+|X_2|}|X_P(v_i)\cap X_D(v_i)|$，$X_P(u_i)$和 $X_D(u_i)$分别表示新增数据集 V 中与 $u_i(u_i\in U)$满足相容关系的数据对象集，$X_P(v_i)$和 $X_D(v_i)$分别表示 U 中与 $v_i(v_i\in V)$满足相容关系的数据对象集。

证明：增加新的数据对象集后，新论域上相容类的变化与定理 1 证明中的分析类似，这里不再作具体分析。令 $m=|U-X_1\cup Y_1|$且 $m_1=|U-X_2\cup Y_2|$，论域 $U\cup V$ 上的信息熵分析如下：

$$\begin{aligned}
&E_{U\cup V}(D|P)\\
&=\frac{1}{|U\cup V|^2}\Big(\sum_{i=1}^{m}(|S_P(u_i)|-|S_P(u_i)\cap S_D(u_i)|)+\sum_{i=m}^{|U-Y_1|}(|S_P'(u_i)|-|S_P'(u_i)\cap S_D'(u_i)|)+\sum_{i=|U-Y_1|}^{m+|X_1|}(|S_P'(u_i)|-|S_P'(u_i)\cap S_D'(u_i)|)+\sum_{i=m+|X_1|}^{|U|}(|S_P'(u_i)|-|S_P'(u_i)\cap S_D'(u_i)|)+\sum_{i=1}^{m'}(|S_P(v_i)|-|S_P(v_i)\cap S_D(v_i)|)+\sum_{i=m'}^{|V-Y_2|}(|S_P'(v_i)|-|S_P'(v_i)\cap S_D'(v_i)|)+\sum_{i=|V-Y_2|}^{m'+|X_2|}(|S_P'(v_i)|-|S_P'(v_i)\cap S_D'(v_i)|)+\sum_{i=m'+|X_2|}^{|V|}(|S_P'(v_i)|-|S_P'(v_i)\cap S_D'(v_i)|)\Big)\\
&=\frac{1}{|U\cup V|^2}\Big(\sum_{i=1}^{m}(|S_P(u_i)|-|S_P(u_i)\cap S_D(u_i)|)+\sum_{i=m}^{|U-Y_1|}(|S_P(u_i)\cup X_P(u_i)|-|(S_P(u_i)\cup X_P(u_i))\cap S_D(u_i)|)+\sum_{i=|U-Y_1|}^{m+|X_1|}(|S_P(u_i)\cup X_P(u_i)|-|(S_P(u_i)\cup X_P(u_i))\cap(S_D(u_i)\cup X_D(u_i))|)+\sum_{i=m+|X_1|}^{|U|}(|S_P(u_i)|-|S_P(u_i)\cap(S_D(u_i)\cup X_D(u_i))|)+\sum_{i=1}^{m'}(|S_P(v_i)|-|S_P(v_i)\cap S_D(v_i)|)+\sum_{i=m'}^{|V-Y_2|}(|S_P(v_i)\cup X_P(v_i)|-|(S_P(v_i)\cup X_P(v_i))\cap S_D(v_i)|)+\sum_{i=|V-Y_2|}^{m'+|X_2|}(|S_P(v_i)\cup X_P(v_i)|-|(S_P(v_i)\cup X_P(v_i))\cap(S_D(v_i)\cup X_D(v_i))|)+\sum_{i=m'+|X_2|}^{|V|}(|S_P(v_i)|-|S_P(v_i)\cap(S_D(v_i)\cup X_D(v_i))|)\Big)\\
&=\frac{1}{|U\cup V|^2}\Big(\sum_{i=1}^{|U|}(|S_P(u_i)|-|S_P(u_i)\cap S_D(u_i)|)+\sum_{i=1}^{|V|}(|S_P(v_i)|-|S_P(v_i)\cap S_D(v_i)|)+\sum_{i=m}^{m+|X_1|}|X_P(u_i)|+\sum_{i=m'}^{m'+|X_2|}|X_P(v_i)|-\sum_{i=|U-Y_1|}^{m+|X_1|}|X_P(u_i)\cap X_D(u_i)|-\sum_{i=|V-Y_2|}^{m'+|X_2|}|X_P(v_i)\cap X_D(v_i)|\Big)
\end{aligned}$$

令 $\Delta=\sum_{i=m}^{m+|X_1|}|X_P(u_i)|+\sum_{i=m'}^{m'+|X_2|}|X_P(v_i)|-\sum_{i=|U-Y_1|}^{m+|X_1|}|X_P(u_i)\cap X_D(u_i)|-\sum_{i=|V-Y_2|}^{m'+|X_2|}|X_P(v_i)\cap X_D(v_i)|$，则有 $E_{U\cup V}(D|P)=\frac{1}{|U\cup V|^2}(|U|^2E_U(D|P)+|V|^2E_V(D|P)+\Delta)$。

由定理 2 可得，当有一组新数据增加到给定的含有缺失数据的数据集中后，通过分析与新增数据集 V 满足相容关系的数据集以及新增数据集上的信息熵，即可计算得到新论域上的互补信息熵，从而避免了在新数据集上重新求解熵的重复计算。

4 增量式特征选择算法

基于上述定理中对互补信息熵增量式更新机制的分析，本节分别介绍面向含有缺失数据数据集的单增量式粗糙特征选择算法和组增量式算法。

4.1 特征重要度度量

定义 2 令 $S=(U,C\cup D)$是一个非完备决策表，$P\subseteq C$，任意属性 $a\in P$ 的特征重要度定义为

$$SIG_{in}(a,P,D)=E(D|P-\{a\})-E(D|P)$$

任意属性 $a\in C-P$ 的特征重要度定义为

$$SIG_{out}(a,P,D)=E(D|P)-E(D|P\cup\{a\})$$

上述定义中的 $SIG_{in}(a,P,D)$称为内部重要度，通常用于删除特征子集中的冗余特征；$SIG_{out}(a,P,D)$称为外部重要度，用于向当前的特征子集中添加新的特征。

4.2 面向含有缺失数据数据集的增量式粗糙特征选择算法

基于互补信息熵增量式更新机制的分析和特征重要度的定义，本小节将介绍增量式的粗糙特征选择算法，具体算法步骤如下。

算法 1 一种面向含有缺失数据数据集的增量式粗糙特征选择算法(IFSA)

输入：非完备决策表 S=(U,C∪D)，论域 U 上的特征选择结果 R_U，新增数据对象 x；

输出：论域 U∪{x}上的特征选择结果 $R_{U\cup\{x\}}$。

步骤 1 B←R_U，计算 X 和 Y：∀u∈U，如果 x 与 u 在 C 上满足相容关系，则 X=X∪{u}，如果 x 与 u 在 D 上满足相容关系，则 Y=Y∪{u}；

步骤 2 计算 X′：∀u∈U，如果 x 与 u 在 B 上满足相容关系，则有 X′=X′∪{u}；

步骤 3 while $E_{U\cup\{x\}}(D|B)\neq E_{U\cup\{x\}}(D|C)$ do

{ ∀a∈C−B，计算其重要度 $SIG_{out}(a,B,D)$；

选择重要度最大的特征 $a_0=\max\{SIG_{out}(a,B,D)\}$，a∈C−B；

B←B∪{a_0}；

}

步骤 4　∀a∈B执行

{ 计算其重要度 $SIG_{in}(a,B,D)$;

如果 $SIG_{in}(a,B,D)=0$,则 $B\leftarrow B-\{a\}$;

}

步骤 5　$R_{U\cup\{x\}}\leftarrow B$,返回 $R_{U\cup\{x\}}$,算法结束。

在同一时刻向给定的含有缺失数据的数据集中增加一组新数据时,为进一步提高计算效率,降低计算耗时,下面介绍组增量式的粗糙特征选择算法。

算法 2　一种面向含有缺失数据数据集的组增量式粗糙特征选择算法(GIFSA)

输入:非完备决策表 $S=(U,C\cup D)$,论域 U 上的特征选择结果 R_U,新增数据对象集 V;

输出:论域 $U\cup V$ 上的特征选择结果 $R_{U\cup V}$。

步骤 1　$B\leftarrow R_U$,计算 X_1,Y_1,X_2 和 Y_2:$\forall u\in U,v\in V$,如果 v 与 u 在 C 上满足相容关系,则 $X_1=X_1\cup\{u\},X_2=X_2\cup\{v\}$,如果 v 与 u 在 D 上满足相容关系,则 $Y_1=Y_1\cup\{u\},Y_2=Y_2\cup\{v\}$;

步骤 2　计算 X_1'和 X_2':$\forall u\in U,v\in V$,如果 v 与 u 在 B 上满足相容关系,则有 $X_1'=X_1'\cup\{u\},X_2'=X_2'\cup\{v\}$;

步骤 3　while $E_{U\cup V}(D|B)\neq E_{U\cup V}(D|C)$ do

{ $\forall a\in C-B$,计算其重要度 $SIG_{out}(a,B,D)$;

选择重要度最大的特征 $a_0=\max\{SIG_{out}(a,B,D)\},a\in C-B$;

$B\leftarrow B\cup\{a_0\}$;

}

步骤 4　∀a∈B执行

{ 计算其重要度 $SIG_{in}(a,B,D)$;

如果 $SIG_{in}(a,B,D)=0$,则 $B\leftarrow B-\{a\}$;

}

步骤 5　$R_{U\cup V}\leftarrow B$,返回 $R_{U\cup V}$,算法结束。

算法 IFSA 的计算时间复杂度分析:根据定理 1,当有 1 个新的数据对象增加到给定数据集中后,增量式求解信息熵的时间复杂度是 $O(|U||C|+|X||Y||C|)$,则有算法步骤 1—步骤 3 的时间复杂度是 $O(|U||C|^2+|X||Y||C|^2)$;步骤 4 的时间复杂度是 $O(|U||C||B|+|X||Y||C||B|)$。所以算法总的时间复杂度是 $O(|U||C|^2+|X||Y||C|^2)$。

算法 GIFSA 的计算时间复杂度分析:根据定理 2,当有 1 个新的数据对象增加到给定数据集中后,增量式求解信息熵的时间复杂度是 $O(|V||U||C|)$,则有算法步骤 1—步骤 3 的时间复杂度是 $O(|V||U||C|^2)$;步骤 4 的时间复杂度是 $O(|V||U||C||B|)$。所以算法总的时间复杂度是 $O(|V||U||C|^2)$。

根据上述分析,使用算法 IFSA 处理一组新增数据对象 V 的时间复杂度是 $O(|V||U||C|^2+|V||X||Y||C|^2)$。与组增量算法 GIFSA 比较,随着新增数据集 $|V|$ 的规模增大,$|V||X||Y||C|^2$ 的值也会明显增大。因此,当有大量的新数据成批地增加到数据集中时,相比单个数据的增量算法,组增量特征选择算法的计算效率会明显提高。

5　实验及分析

为验证组增量算法 GIFSA 的高效性和可行性,本文选取了 4 组含有缺失数据的 UCI 数据集(见表 1)进行测试。运行实验程序的个人计算机的配置为:CPU Inter(R) Core(TM)2 Quad CPU Q9400,2.66GHz,内存为 4.00GB,操作系统是 Windows 7。程序开发平台是 Microsoft Visual Studio 2005,编程语言为 C#。本节中实验分两部分,5.1 节中通过比较分别由组增量算法 GIFSA 与传统的非增量特征选择算法计算得到的特征选择结果来验证算法 GIFSA 的可行性,5.2节中则通过比较组增量算法 GIFSA 和单增量算法 IFSA 的计算时间来验证算法 GIFSA 的高效性。具体的实验数据介绍如表 1 所列。

表 1　数据集描述

	数据集	样本数	特征数	类别数
1	ticdata2000	5822	85	2
2	mushroom	8124	22	2
3	adult	48842	14	2
4	shuttle	58000	9	7

5.1　可行性实验分析

对于表 1 中每组数据集,选取 60%的数据作为基础数据集,剩余 40%为增量数据集。当增量数据集增加到基础数据集中后,本节分别使用本文中提出的组增量式算法 GIFSA 与基于互补熵的非增量算法来求解新数据集上的特征选择结果,并选取了机器学习中两个常见分类器朴素贝叶斯(NBC)和决策树(C4.5)来比较特征选择结果的分类精度。实验结果见表 2,其中 N 表示特征选择结果中特征的个数,CFS 表示基于互补熵的非增量粗糙特征选择算法。

表 2　特征选择结果

数据集	CFS				GIFSA			
	N	NBC	C4.5	时间/s	N	NBC	C4.5	时间/s
ticdata2000	26	0.9002±0.1291	0.9402±0.1124	1785.53	24	0.9016±0.1284	0.9402±0.1124	179.33
mushroom	7	0.9825±0.0261	1.0000±0.0000	175.01	5	0.9883±0.0208	0.9980±0.0034	59.98
adult	11	0.7974±0.2225	0.8541±0.2077	23847.44	10	0.8095±0.2200	0.8533±0.2091	685.38
shuttle	4	0.8155±0.0667	0.9972±0.0010	7414.73	4	0.8155±0.0667	0.9972±0.0010	321.16

由表 2 中的实验结果可得,由组增量算法 GIFSA 求解得到的特征子集在分类器 NBC 和 C4.5 上的分类精度与由传统的非增量算法求解得到的特征子集的分类精度是非常接近的,即算法 GIFSA 可找到一个有效的特征子集。而计算时间的比较则验证了 GIFSA 的高效性,尤其对一些规模较大的数据集,GIFSA 的高效性更加明显。因此,针对动态增加的数据集,与非增量算法比较,组增量算法 GIFSA 可高效地求解到一个有效的特征子集。

5.2　高效性实验分析

为进一步验证组增量算法 *GIFSA* 的高效性,对表 1 中的每组数据集,选取 60%的数据作为基础数据集,剩余的 40%平均分成 5 份,记为 $x_i(i=1,2,\cdots,5)$,令 $X_i=\bigcup_{j=1}^{i}x_i(i=1,2,\cdots,5)$为 5 个增量数据集。本节主要比较将上述 5 个增量数据集依次增加到基础数据集中,并分别使用算法 IFSA 和 GIFSA 求解特征选择结果的计算时间。实验结果见图 1—图 4,其中 y 轴表示求解特征选择结果的计算时间,x 轴表示增

加的不同规模的数据集，x 轴上的值 1,2,3,4 和 5 分别表示增加 X_1,X_2,X_3,X_4 和 X_5 到基础数据集中。

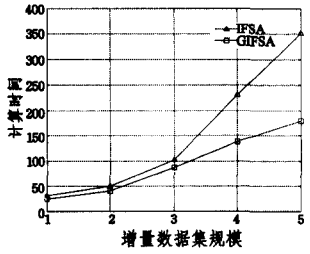

图 1 IFSA 和 GIFSA 基于数据集 ticdata2000 的计算时间

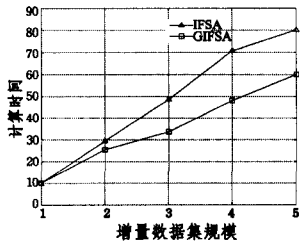

图 2 IFSA 和 GIFSA 基于数据集 mushroom 的计算时间

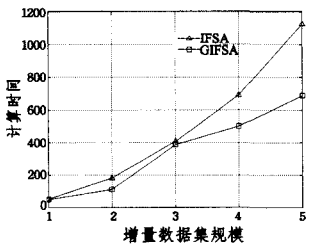

图 3 IFSA 和 GIFSA 基于数据集 adult 的计算时间

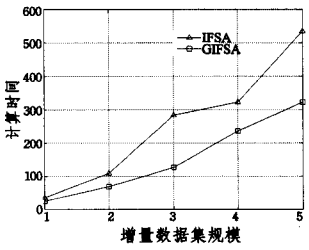

图 4 IFSA 和 GIFSA 基于数据集 shuttle 的计算时间

图 1—图 4 的实验结果表明，当 1 组新数据增加到基础数据集中时，组增量算法 GIFSA 的计算效率明显优于单增量算法 IFSA；而且随着新增数据集规模的不断增大，组增量算法 GIFSA 的高效性越来越明显。因此，本节的实验结果进一步证明了组增量算法的高效性。

结束语 面向动态数据集的数据挖掘已逐渐成为智能信息领域中的一个热点研究问题。本文以含有缺失数据的动态数据集为研究对象，基于粗糙集理论，发展了特征选择的组增量算法。实验分析进一步证明了新算法可高效地一次处理一批新增的数据集。新算法为高效地处理动态数据集提供了一定的方法和基础，也为在线信息的有效处理提供了新的视角和研究途径。

参 考 文 献

[1] Blum A L, Langley P. Selection of relevant features and examples in machine learning [J]. Artificial Intelligence, 1997, 97: 245-271

[2] Liu H, Yu L. Toward integrating feature selection algorithms for classification and clustering [J]. IEEE Transaction on Knowledge and Data Engineering, 2005, 17(4): 491-502

[3] Jain A, Zongker D. Feature selection: evaluation, application, and small sample performance [J]. IEEE Transaction on Pattern Analysis and Machine Intelligence, 1997, 19(2): 153-158

[4] Liang J Y, Wang F, Dang C Y, et al. An efficient rough feature selection algorithm with a multi-granulation view [J]. International Journal of Approximate Reasoning, 2012, 53: 912-926

[5] 王锋，梁吉业，钱宇华. 序信息系统的启发式属性约简算法[J]. 计算机科学，2010，37(3)：258-260，278
Wang Feng, Liang Ji-ye, Qian Yu-hua. Heuristic attribute reduction algorithm to ordered information system[J]. Computer Science, 2010, 37(3): 258-260, 278

[6] 徐章艳，刘作鹏，杨炳儒，等. 一个复杂度为 max(O(|C||U|), O(|C|²|U/C|)) 的快速属性约简算法[J]. 计算机学报，2006，29(3)：391-399
Xu Zhang-yan, Liu Zuo-peng, Yang Bing-ru, et al. A quick attribute reduction algorithm with complexity of max (O(|C||U|), O(|C|²|U/C|))[J]. Chinese Journal of Computers, 2006, 29(3): 391-399

[7] Hu Q H, Xie Z X, Yu D R. Hybrid attribute reduction based on a novel fuzzy-rough model and information granulation [J]. Pattern Recognition, 2007, 40: 3509-3521

[8] Wang F, Liang J Y, Qian Y H. Attribute reduction: a dimension incremental strategy [J]. Knowledge-Based Systems, 2013, 39: 95-108

[9] Wang F, Liang J Y, Dang C Y. Attribute reduction for dynamic data sets [J]. Applied Soft Computing, 2013, 13: 676-689

[10] Liu D, Li T R, Ruan D, et al. An incremental approach for inducing knowledge from dynamic information systems [J]. Fundamenta Informaticae, 2009, 94: 245- 260

[11] 刘宗田. 属性最小约简的增量式算法[J]. 电子学报，1999，27(11)：96-98
Liu Zong-tian. An incremental arithmetic for the smallest reduction of attributes[J]. Acta Electronica Sinica, 1999, 27(11): 96-98

[12] Orlowska M, Orlowski M. Maintenance of knowledge in dynamic information systems [M]// Slowinski R, ed. Intelligent Decision Support: Handbook of Applications and Advances of the Rough Set Theory. Kluwer Academic Publishers, Dordrecht, 1992: 315-330

[13] 杨明. 一种基于改进差别矩阵的核增量式更新算法[J]. 计算机学报，2006，29(3)：407-413
Yang Ming. An incremental updating algorithm for core based on improved discernibility matrix[J]. Chinese Journal of Computers, 2006, 29(3): 407-413

[14] 杨明. 一种基于改进差别矩阵的属性约简增量式更新算法[J]. 计算机学报，2007，30(5)：815-822
Yang Ming. An incremental updating algorithm for attribute reduction based on improved discernibility matrix [J]. Chinese Journal of Computers, 2007, 30(5): 815-822

[15] Hu F, Wang G Y, Huang H, et al. Incremental attribute reduction based on elementary sets [C]// Proceedings of the 10th International Conference on Rough Sets, Fuzzy Sets, Data Mining and Granular Computing. Regina, Canada, 2005: 185-193

[16] 梁吉业，魏巍，钱宇华. 一种基于条件熵的增量核求解方法[J]. 系统工程理论与实践，2008，28(4)：81-89
Liang Ji-ye, Wei Wei, Qian Yu-hua. An incremental approach to computing of a core based on conditional entropy[J]. Chinese Journal of System Engineering Theory and Practice, 2008, 28(4): 81-89

[17] 刘薇，梁吉业，魏巍，等. 一种基于条件熵的增量式属性约简求解算法[J]. 计算机科学，2011，38(1)：229-231，239
Liu Wei, Liang Ji-ye, Wei Wei, et al. An incremental approach to computing of attribute reduction based on conditional entropy [J]. Computer Science, 2011, 38(1): 229-231, 239

[18] Liang J Y, Wang F, Dang C Y, et al. A group incremental approach to feature selection applying rough set technique [J]. IEEE Transactions on Knowledge and Data Engineering, 2014, 26(2): 294-308

[19] Kryszkiewicz M. Rough set approach to incomplete information systems [J]. Information Sciences, 1998, 112: 39-49

[20] Slowinskir R, Vsnderprooten D. A generalized definition of rough approximations based on similarity [J]. IEEE Transactions on Data and Knowledge Engineering, 2000, 12(2): 331-336

[21] Liang J Y, Xu Z B. The algorithm on knowledge reduction in incomplete information systems [J]. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2002, 10(1): 95-103

[22] 黄兵，周献中，张蓉蓉. 基于信息量的不完备信息系统属性约简[J]. 系统工程理论与实践，2005，25(4)：55-60
Huang Bing, Zhou Xian-zhong, Zhang Rong-rong. Attribute reduction based on information quantity under incomplete information systems[J]. Chinese Journal of System Engineering Theory and Practice, 2005, 25(4): 55-60

[23] 钱宇华，梁吉业，王锋. 面向非完备决策表的正向近似特征选择加速算法[J]. 计算机学报，2011，34(3)：435-442
Qian Yu-hua, Liang Ji-ye, Wang Feng. A positive approximation based accelerated algorithm to feature selection from incomplete decision tables [J]. Chinese Journal of Computers, 2011, 34(3): 435-442

[24] Liang J Y, Shi Z Z, Li D Y, et al. The information entropy, rough entropy and knowledge granulation in incomplete information systems [J]. International Journal of General Systems, 2006, 34(1): 641-654

[25] Li T R, Ruan D, Geert 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

[26] Pawlak Z, Skowron A. Rudiments of rough sets [J]. Information Sciences, 2007, 177(1): 3-27

(上接第 257 页)

在两个数据集上的实验结果表明，WNNM 模型应用在矩阵填充上得到了较好的效果，在速度上和精度上有所提高，但是在矩阵规模大的情况下可能优势不明显。

结束语 矩阵填充是目前重要的数据分析工具，在图像处理、文本分析等方面都有重要应用，是近些年的研究热点。本文总结了矩阵填充问题的研究现状，针对原有矩阵填充模型，提出了一种加权核范数最小化模型，提高了核范数的灵活度。在真实数据集上进行的实验表明，该模型可以使矩阵填充在速度和精度上得到一定程度的提升。

同矩阵填充问题等价的低秩矩阵分解、鲁棒性主成分分析问题等将应用于越来越多的方向，所以解决这类问题的算法仍需要不断优化，更要在并行和分布式上做考虑。未来的工作将进一步完善大规模数据上的矩阵填充算法，使其在动态变化的矩阵上也能得到较好的应用。

参 考 文 献

[1] 马宏伟，张光卫，李鹏. 协同过滤推荐算法综述[J]. 小型微型计算机系统，2009，30(7)：1282-1288
Ma Hong-wei, Zhang Guang-wei, Li Peng. Survey of Collaborative Filtering Algorithms[J]. Journal of Chinese Computer Systems, 2009, 30(7): 1282-1288

[2] 陈敏铭. 矩阵恢复与重建[D]. 北京：中国科学院计算技术研究所，2010
Chen Min-ming. Algorithms and Implementation of Matrix Reconstruction[D]. Beijing: Chinese Academy of Sciences, 2010

[3] Zhang F, Chang H. A collaborative filtering algorithm embedded BP network to ameliorate sparsity issue[C]// Proceedings of 2005 International Conference on Machine Learning and Cybernetics, 2005. IEEE, 2005: 1839-1844

[4] Jung K Y, Hwang H J, Kang U G. Constructing full matrix through naive Bayesian for collaborative filtering[M]// Computational Intelligence. Springer Berlin Heidelberg, 2006: 1210-1215

[5] Candès E J, Recht B. Exact matrix completion via convex optimization[J]. Foundations of Computational mathematics, 2009, 9(6): 717-772

[6] Toh K C, Yun S. An accelerated proximal gradient algorithm for nuclear norm regularized linear least squares problems[J]. Pacific Journal of Optimization, 2010, 6(15): 615-640

[7] Lin Zhou-chen, Chen Min-ming, Ma Yi. The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices[J]. arXiv preprint arXiv: 1009. 5055, 2010

[8] Tao M, Yuan X. Recovering low-rank and sparse components of matrices from incomplete and noisy observations [J]. SIAM Journal on Optimization, 2011, 21(1): 57-81

[9] Yuan X, Yang J. Sparse and low-rank matrix decomposition via alternating direction methods[J]. Preprint, 2009, 9(1): 1-16

[10] Yang J, Yuan X. Linearized augmented Lagrangian and alternating direction methods for nuclear norm minimization[J]. Mathematics of Computation, 2013, 82(281): 301-329

[11] Lin Z, Liu R, Su Z. Linearized Alternating Direction Method with Adaptive Penalty for Low-Rank Representation[J]// arXiv: 1109. 0367, 2011

[12] Cai J F, Candès E J, Shen Z. A singular value thresholding algorithm for matrix completion[J]. SIAM Journal on Optimization, 2010, 20(4): 1956-1982

[13] Gu S, Zhang L, Zuo W, et al. Weighted Nuclear Norm Minimization with Application to Image Denoising[C]// 2014 IEEE Conference on Computer Vision and Pattern Recognition. 2014: 2862-2869

[14] Xie Q, Meng D, Gu S, et al. On the Optimal Solution of Weighted Nuclear Norm Minimization[J]. arXiv preprint arXiv: 1405. 6012, 2014