基于标识特征的信息系统特征选取*)

李国和

(中国石油大学(北京)计算机科学与技术系　北京 102249)

摘　要　通过引入标识特征集,把信息系统分解为独立的信息子系统。各个信息子系统逐一转变为类扩张矩阵和浓缩类扩张矩阵。以每个类扩张矩阵的统计信息为启发式信息,逐步完成信息子系统的特征选取和整合,最终形成整个信息系统的特征选取。这种特征选取方法具有高效和较优的特点。

关键词　信息系统,信息子系统,特征选取,启发式信息,扩张矩阵

Feature Selection of Information System Based on Specified Features

LI Guo-He

(Dept. of Computer Science and Technology, China University of Petroleum, Beijing 102249)

Abstract　Information system is divided into information subsystems by the specified features, and then each subsystem is turned into Condensed Similar Extension-Matrix(CSEM). By means of the statistical values as heuristic information and CSEM, feature subset of each information subsystem is obtained in turn, and then gradually integrated into a feature subset of information system. The approach of feature selection for information system has the advantage of efficiency and feature selection is often close to minimum.

Keywords　Information system, Information subsystem, Feature selection, Heuristic information, Extension matrix

1　引言

冗余数据不仅降低学习速度,而且也降低知识模型的可理解性,从而导致应用知识模型进行预测和识别决策的准确性[1]。通过特征选取找出能够描述某一应用领域的特征子集,使得该特征子集与原有的特征集具有相同的信息量,达到消除冗余的目的。目前,国内外有多种特征选取的方法[1~9],其中大多是针对决策表中条件特征进行特征选取,只有少数是针对信息系统进行特征选取[4~6]。所有这些特征选取方法把信息系统或决策表作为不可分割的整体进行处理。当信息系统较大时,必然受到有限计算机资源的限制。由于特征选取往往并不唯一,而且是 NP-Hard 难题[2,9],因此,主要采用最优化方法求解一个最优或较优的特征选取。针对较大信息系统,本文实现了高效的较优特征选取。

2　相关基本概念

定义 1　对于信息系统 $K(U,A,V)$,存在特征子集 $Fs\subseteq A$,且 U 的两个划分 U/A、U/Fs 相等(即 $U/A=U/Fs$),称特征子集 Fs 是 K 的特征选取。

如果 $Fs\subseteq A$ 是 K 的特征选取,对于任一 $r\in A$,那么 $Fs\cup\{r\}$ 仍然为 K 的特征选取。如果 Fs 是 K 的特征选取,且对任一 $r\in Fs$,$U/Fs\neq U/(Fs-\{r\})$,称 Fs 是 K 的最小特征选取[6]。

定义 2　$S_K(U\times U,A,\{\{0,1\}\})$为信息系统 K 导出的信息系统,对于 $\forall a\in A$,$\forall\ (u,v)\in U\times U$,特征取值为:

$$a((u,v))=\begin{cases}1,\text{当 } a(u)\neq a(v)\\0,\text{当 } a(u)=a(v)\end{cases}$$

称 S_K 为信息系统 K 的类扩张矩阵。

定义 3　$S_K(U\times U,A,\{\{0,1\}\})$为信息系统 K 的类扩张矩阵,对于特征 $a\in A$,集合 $DisSet_a=\{(u,v)\mid\forall\ (u,v)\in U\times U$,且 $a((u,v))=1\}$,称 $DisSet_a$ 为类扩张矩阵 S_K 中基于 a 的对象可识别集。

实际上,对象可识别集 $DisSet_a$ 表明了 a 可以区分的对象。

定理 1[12]　$S_K(U\times U,A,\{\{0,1\}\})$为信息系统 K 的类扩张矩阵。特征子集 Fs 为信息系统 K 的特征选取,当且仅当 $\bigcup_{\forall a\in Fs}DisSet_a=U\times U$(其中为 $DisSet_a$ 基于特征 a 的对象可识别集)。

这一定理表明可以通过类扩张矩阵 S_K 求解信息系统 K 的最小特征子集。

定义 4　$S_K(U\times U,A,\{\{0,1\}\})$为信息系统 $K(U,A,V)$ 的类扩张矩阵,信息系统 $Red(S_K)(U'\times U',A,\{\{0,1\}\})$,其中 $U'\times U'=\{(u,v)\mid\forall\ (u,v),(p,q)\in U\times U$,且 $m_{u,v},m_{p,q},m_{u,v}\times m_{p,q}\}$,称 $Red(S_K)$为类扩张矩阵 S_K 的浓缩类扩张矩阵。

根据上述内容,基于类扩张矩阵的特征选取[12](Efficient and Optimal Feature Selection,EOFS)包括如下两个子过程:

(1)FormCSEM(K,$Red(S_K)$,$TotalInfo$)

该子过程实现信息系统 K 形成浓缩类扩张矩阵 $Red(S_K)$和统计信息 TotalInfo。

(2)FormFeatureSubSet($Red(S_K)$,$TotalInfo$,Fs)

该子过程根据定理 1,在浓缩类扩张矩阵 $Red(S_K)$中进行特征选取。

尽管 EOFS 执行效率优于基于粗糙集理论的特征选取方

*)本研究得到国家自然科学基金资助项目(60473125)、中国石油(CNPC)石油科技中青年创新基金资助项目(05E7013)的资助。李国和　博士,教授,研究方向:人工智能、知识发现、数据库技术、计算机图形等。

法[12]，但仍只适用于信息系统不太大的情况。当信息系统很大时，构造 $Red(S_K)$ 成为时间和空间资源的主要开销，为此进一步提出了基于标识特征的特征选取（Feature Selection Based on Specified Features，FSSF）。

3 基于标识特征的特征选取

为了建立基于标识特征的特征选取 FSSF，先给出定义和一些性质。

3.1 FSSF 相关基本概念

定义 5 $K(U,A,V)$，$K_1(U_1,A,V)$，$K_2(U_2,A,V)$ 为信息系统，且 $\{U_1,U_2\}$ 是 U 的一个划分，称 K_1、K_2 为 K 的完全分解信息系统。

很显然有性质：若特征子集 $R\subseteq A$ 是 K 的特征选取，则 R 也为 K_1、K_2 的特征选取。若存在特征子集 $ID\subset A$，$ID\neq\phi$，且 $U_1/ID=\{U_1\}$，$U_2/ID=\{U_2\}$，R_1、R_2 分别为 K_1、K_2 的特征选取，则 $R=ID\cup R_1\cup R_2$ 是 K 的特征选取。把 ID 称为 K 的完全分解信息系统 K_1、K_2 的标识特征。若 R 为 K_1，K_2 的特征选取，且 $ID\subseteq R$，则 R 也是 K 的特征选取。这些性质可以推广到：以标识特征集 ID 为等价关系，得到完全分解信息系统 $K_i(U_i,A,V)$（其中 $i=1,2,\cdots,|U/ID|$，$U_i\in U/ID$）的情形。

3.2 FSSF 过程

根据上述思想，结合浓缩类扩张矩阵和差异统计启发式信息，FSSF 实现一种高效特征选取：

```
Sub FSSF(K,ID,Fs) '输入 K、ID,输出 Fs
    Ks={Ki={Ui, A, V}| Ui∈U/ID}
    Ks=Ks-{K1}, K1∈Ks '第一个信息子系统
    FormCSEM(K1,Red(SK1),TotalInfo1)
              '第一个信息子系统特征选取
    FormFeatureSubSet(Red(SK1),TotalInfo1,Fs)
    Fs=ID∪Fs '整合特征选取
    For Each Ki∈Ks '下一个信息子系统
      FormCSEM(Ki,Red(SKi),TotalInfoi)
            '浓缩类扩张矩阵和统计信息
      Red(SKi)=Red(SKi)- ∪ DisSeta
                        a∈Fs
      For Each a∈Fs, TotalInfoi(a)=0
      FormFeatureSubSet(Red(SKi),TotalInfoi,Fsi) '特征子集
      Fs=Fs∪Fsi'整合特征选取
    End For
End Sub
```

FSSF 对于第一个信息子系统 K_1 采用 EOFS 进行特征子集 Fs 的求解，对于余下的信息子系统 $K_i(i>1)$ 采用构造浓缩扩张矩阵 $Red(S_{Ki})$ 和统计信息 $TotalInfo_i$，利用已求得的特征子集 Fs 对 $Red(S_{Ki})$ 进行初步减小数据空间，并修正相应的统计信息为 0，最后根据当前的 $Red(S_{Ki})$ 和 $TotalInfo_i$ 完成特征选取。

4 算法分析与实验结果

本文中，空间开销指的是特征值的个数。时间开销指的是特征值的比较次数。

4.1 FSSF 分析

根据 FSSF，信息系统 $K(U,A,V)$ 完全分解为一系列信息子系统 $Ks=\{K_i=\{U_i,A,V\}|U_i\in U/ID\}$。令 $m=|U/ID|$。

对于 U_n（其中 $n>1$）的特征子集为 Fs_n，$Fs_n=Fs_{n1}\cup Fs_{n2}$，$Fs_{n1}=Fs_n-Fs$，$Fs_{n2}=Fs\cap Fs_n$（Fs 是进行第 $1,2,\cdots,n-1$ 次特征选取的特征子集），对于任一特征 $f_j\in Fs_n$ 的概率分布为 $\{P_j(U_n)|j=\{1,2,\cdots,|Fs_n|\}\}$，求解 Fs_{n1}、Fs_{n2} 特征子集的平均时间开销分别为 $Cost_{T,Axe01}(U_n)$，$Cost_{T,Axe02}(U_n)$：

$$Cost_{T,Axe01}(U_n)=\frac{|U_n|\times(|U_n|-1)}{2}\times(|Fs_{n1}|-\sum_{i=1}^{|Fs_{n1}|-1}\sum_{j=1}^{i}P_j(U_n)) \quad (1)$$

$$Cost_{T,Axe02}(U_n)=\frac{|U_n|\times(|U_n|-1)}{2}\times(|Fs_{n2}|-\sum_{i=1}^{|Fs_{n2}|-1}\sum_{j=1}^{i}P_j(U_n)) \quad (2)$$

而求解 Fs_n 特征子集的平均时间开销 $Cost_{T,Axe0}(U_n)$ 为：

$$Cost_{T,Axe0}(U_n)=Cost_{T,Axe01}(U_n)+Cost_{T,Axe02}(U_n)$$
$$=\frac{|U_n|\times(|U_n|-1)}{2}\times(|Fs_n|-\sum_{i=1}^{|Fs_{n1}|-1}\sum_{j=1}^{i}P_j(U_n)-\sum_{i=1}^{|Fs_{n2}|-1}\sum_{j=1}^{i}P_j(U_n)) \quad (3)$$

求解 Fs 特征子集的平均时间开销 $Cost_{T,Axe0}$ 为：

$$Cost_{T,Axe0}=Cost_{T,Axe0}(U_1)+\sum_{n=2}^{m}Cost_{T,Axe0}(U_n) \quad (4)$$

如果每个 K_i 的规模一样，那么求解 Fs 特征子集的平均时间开销 $Cost_{T,Axe0}$ 为：

$$Cost_{T,Axe0}=\frac{|U|\times(|U|-m)}{2m^2}\times fun(\{P\}) \quad (5)$$

其中，$fun(\{P\})=\sum_{n=1}^{m}|Fs_n|-\sum_{i=1}^{Fs_1-1}\sum_{j=1}^{i}P_j(U_1)-\sum_{i=1}^{m}(\sum_{i=1}^{Fs_{n1}}\sum_{j=1}^{i}P_j(U_n)+\sum_{i=1}^{Fs_{n2}}\sum_{j=1}^{i}P_j(U_n))$。

由于 $Fs_n\subseteq Fs$，$Cost_{T,Axe0}\leqslant\frac{|U|\times(|U|-m)}{2m}\times(|Fs|-\frac{fun(\{P\})}{m})$。随着 m 的增大，对概率分布进行从大到小的排序，$fun(\{P\})$ 也在增大，$\frac{fun(\{P\})}{m}$ 不一定减小，但 $\frac{|U|\times(|U|-m)}{2m}$ 却明显下降，表明效率快速提高。

4.2 FSSF 实验

下面进行三个实验：随机产生 100 个基础数据库，字段数为 10，记录数为 200。每 10 个数据库一组，分别进行特征选取，统计每一组求解平均时间和特征子集长度。后两个实验中，数据库在基础数据库基础上进行一些变化。

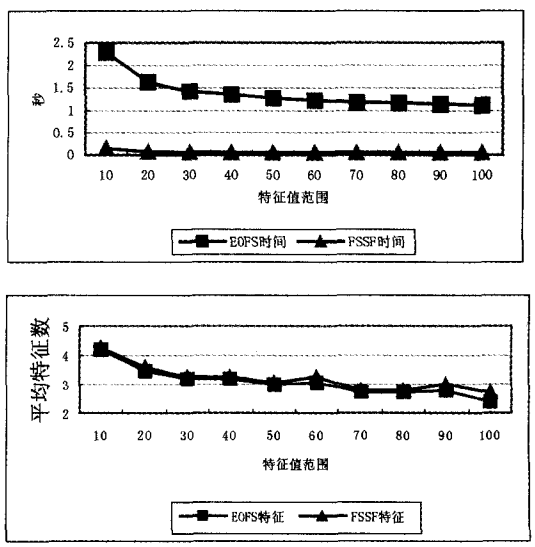
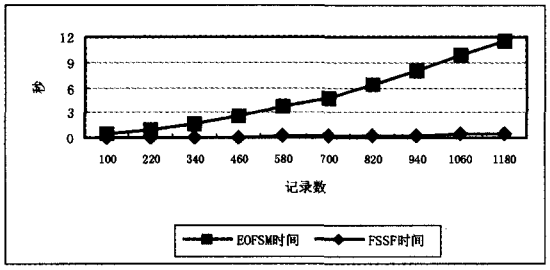


图 1 特征值变化，记录数和字段数不变

实验 1 同一组内数据库字段取值范围一样，不同组数据库字段取值范围不同，并按 10 递增，也就是取值范围大，对象的差异可能就大，信息子系统可能就小，但个数多。从图 1

中可以看出，EOFS特征选取的时间大于FSSF的特征选取时间，但两种特征子集的大小基本相同。

实验2 同一组内数据库记录数一样，不同组数据库记录数不同，并按120递增，也就是每一组内信息子系统大小基本一样，不同组之间信息子系统大小不一样。从图2中可以看出，EOFS特征选取的时间大于FSSF的特征选取时间，而且随着记录数的增加，EOFS求解时间快速增长，而FSSF时间增长较为缓慢，但两种特征子集的大小基本相同。



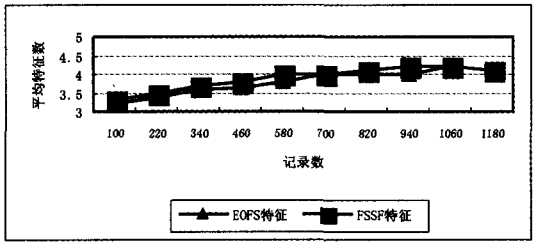
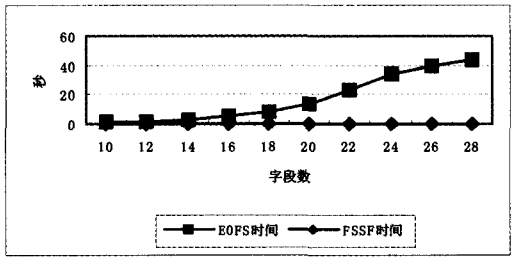


图2 记录数变化，特征值和字段数不变



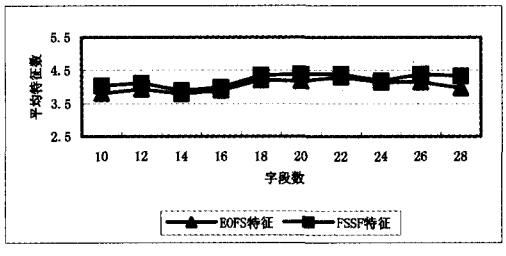


图3 字段数变化，特征值和记录数不变

实验3 同一组内数据库字段数一样，不同组数据库字段数不同，并按2递增，但每一组内信息子系统大小基本相同。从图3中可以看出，EOFS特征选取的时间大于FSSF的特征选取时间，而且随着字段数的增加，EOFS求解时间快速增长，而标FSSF时间增长较为缓慢，但两种特征子集的大小基本相同。

结束语 FSSF不仅具有高效的特点，而且可以克服要求大内存的限制，特别适用于大信息系统的特征选取。通过分析和实验可以看出，FSSF的关键在于：①把信息系统按标识特征进行分解，确保信息子系统较优的特征选取；②在信息子系统中，采用差异统计启发式信息快速减小数据空间；③利用标识特征进行特征子集的综合，确保了特征选取是信息系统的全局较优。尽管FSSF求解的特征选取不一定是最优的，但如果标识特征具有较强的对象分类能力，那么FSSF的特征选取越接近于全局最优的特征选取。因此，在工程应用上具有实用价值。

参考文献

1 戴东亚，郑启伦，胡劲松，陈小航. 一种基于粗糙集的混合特征选取方法[C]. 计算机科学，2001，28(5专刊)：95～97

2 陈彬，洪家荣，王亚东. 最优特征子集选取[J]. 计算机学报，1997，20(2)：133～138

3 朱明，王俊普，蔡庆生. 一种最优特征集的选取算法[J]. 计算机研究与发展，1998，35(9)：803～805

4 Kohavi R，Frasca B. Useful Feature Subsets and Rough Set Reducts[C]. In：3th International Workshop on RoughSets and Soft Computing. 1994

5 李萌，魏长华. 一种基于差异矩阵的属性简约算法[C]. 计算机科学，2002，29(9专刊)：403～406

6 曾黄麟. 粗集理论及其应用[M]. 重庆：重庆大学出版社，1998. 55～70

7 Gasca E，SÃÃnchez J S，Alonso R. Eliminating redundancy and irrelevance using a new MLP-based feature selection method[J]. Pattern Recognition，2006，39(2)：313～315

8 Igor K，Hong S J. Attribute selection for modelling[J]. Future Generation Computer Systems，1994. 121～129

9 Skowron A，Rauszer C. The Discernibility Matrices and Function in Information Systems. In：Slowinski R，ed. Intelligent Decision Support-Handbook of Application and Advances of Rough Set Theory[M]. Kuwler Academic Publisher，1997，13(2-3)：181～195

10 洪家荣. 示例学习的扩张理论[J]. 计算机学报，1991，6(6)：401～410

11 汪培庄，李洪兴. 模糊系统理论与模糊计算机[M]. 北京：科学出版社，1996. 40～52

12 李国和. 基于类扩张矩阵的信息系统特征选取[J]. 计算机工程，2006，32(17)：52～54