**文章编号**：1003-0077(2020)02-0056-07

# 基于 GMM 的文本规则挖掘的粗糙集方法研究

洪壮壮，黄兆华，万仲保，张 薇，高梦茜

（华东交通大学 软件学院，江西 南昌 330013）

**摘 要**：领域文本具有结构复杂、相似性高以及动态变化等特点，且存在着连续型与离散型并存的混合数据，这在一定程度上限制了知识发现方法对文本规则的挖掘效率。针对这一问题，该文提出了基于 GMM 与粗糙集的文本规则挖掘方法。该方法首先根据目标数据的属性类型构造信息表；然后利用高斯混合模型（GMM，Gaussian Mixture Model）聚类算法对连续数据进行聚类划分，依此对数据进行离散化及状态约简，并生成决策表；最后利用粗糙集理论对决策表进行属性约简，通过约简表对决策规则进行提取。实验结果表明：相比于传统的方法，该文方法拥有更高的抽取精度以及较强的属性约简能力，其信息抽取的平均准确率与 $F_1$ 值能够达到 95.0%和 95.7%。

**关键字**：混合数据；规则挖掘；高斯混合模型；粗糙集；属性约简；决策规则

**中图分类号**：TP391 **文献标识码**：A

## Research on Rough Set Method of Text Rule Mining Based on GMM

HONG Zhuangzhuang, HUANG Zhaohua, WAN Zhongbao, ZHANG Wei, GAO Mengxi

(Department of Software Engineering, East China Jiaotong University, Nanchang, Jiangxi 330013, China)

**Abstract**: The domain texts can be characterized by the complex structure, the high similarity and the dynamic change. With a mixture of continuous and discrete types of data, the existing knowledge discovery method is restricted in the mining efficiency of the text rules. To deal with this issue, this paper proposes a text rule mining method based on GMM and Rough Set. Firstly, the method constructs an information table according to the attribute type of the target data; Then, the Gaussian Mixture Model (GMM) clustering algorithm is applied to cluster the continuous data, on which the data is discretized and the state is reduced, and the decision table is generated; Finally, the rough set theory is used to reduce the attributes of decision table, and the decision rules are extracted through the reduction table. The experimental results show that the proposed method has higher precision and stronger attribute reduction ability, achieving an average precision and $F$ score of 95.0% and 95.7%, respectively.

**Keywords**: hybrid data; rule mining; Gaussian Mixture Model; rough set; attribute reduction; decision rule

## 0 引言

随着互联网技术的快速发展，网络上的信息呈现爆炸式增长，大量数据在各个领域的成功应用宣告了大数据时代的来临。与此同时，大数据具有动态性、异构性以及随机性等一系列特征，这些特征均体现了大数据的不确定性。因此，如何从这些不确定性极强的数据中挖掘出有价值的知识信息，已成为一个日渐重要的研究课题。

粗糙集理论[1]是由波兰学者 Z Pawlak 于 1982 年提出的一种处理模糊性与不确定性信息的数学工具，它主要是通过上近似、下近似算子来对不确定性信息进行刻画。粗糙集理论具有不需要先验知识的特点，因此相比于其他处理不确定性知识的方法，其更具有实用性，目前已在数据挖掘[2]、机器学习[3]以及模式识别[4]等领域得到了广泛应用。属性约简是粗糙集理论的重要研究内容之一，其核心思想是在确保已有知识库分类能力不变的情形下，删除数据中的冗余属性，以此达到降低数据规模及简化数据结构的目的，便于问题分类规则的导出，因此属性约简可以视为获取规则过程中最核心的问题。李俊

**收稿日期**：2019-08-05 **定稿日期**：2019-10-28

**基金项目**：国家重点研发计划(2018YFC0831106)；江西省自然科学基金(20122BAB201040)

等[5]利用粗糙集提取了导弹的质量性能评估规则集，提高了导弹质量评估的准确率；张腾飞等[6]根据拓展的优势关系，提出了相对下、上近似约简，从而能够在不完备决策信息系统中提取简化的决策规则；卢娇丽等[7]运用粗糙集对文本特征词向量进行属性约简及规则抽取，利用规则对文本进行分类。然而传统粗糙集理论建立在等价关系的基础上，只适用于离散型数据，而在领域文本中存在着离散型与连续型属性并存的混合数据。因此，如何利用粗糙集理论对领域文本中的混合数据实现规则挖掘是一个难题。

针对上述问题，本文提出了基于 GMM 与粗糙集的领域文本规则的挖掘方法。该方法首先根据目标数据属性构造信息表；然后通过高斯混合模型聚类算法对连续型数据进行离散化处理，同时在此基础上进行状态约简，并生成决策表；最后利用粗糙集理论对决策表进行属性约简，消除冗余属性，得到决策表的属性约简结果，随之导出决策规则。实验结果表明，利用该方法得到的规则能够对领域文本中的混合数据进行很好的抽取。

## 1　相关知识

### 1.1　聚类

聚类是将一组对象划分为若干个簇，使得同一个簇中的对象相似性尽可能高，不同簇之间的对象相似性尽可能低。从机器学习的角度来看，聚类属于无监督学习，其目的是通过对无标签数据进行划分来发现目标数据的结构表示，是知识发现的重要手段。目前常用的聚类算法包括：划分聚类法[8]、网格聚类法[9]、密度聚类法[10]、层次聚类法[11]、模型聚类法[12]等。聚类中常使用欧氏距离作为对象(簇)之间相似度的衡量标准，定义如式(1)所示。

$$S_{xy}=\left\{\sum_{i=1}^{n}(f(x,a_i)-f(y,a_i))^2\right\}^{\frac{1}{2}} \tag{1}$$

其中，$x,y\in U$，$n$ 为对象的维数(即属性数目)，$f(x,a_i)$，$f(y,a_i)$分别为对象 $x$，$y$ 第 $i$ 个维度的值，且 $1\leqslant i\leqslant n$。

### 1.2　粗糙集理论

粗糙集理论是一种处理不确定性、不完整性信息的数学工具。粗糙集理论不需要任何先验知识，它能够直接从具体的问题描述出发，分析数据之间的内在规律，具有极强的定性分析能力。近年来，随着大数据及智能计算技术的飞速发展，对于异构数据的处理需求也不断增加，因而粗糙集理论的特征选择、知识发现具有更重要的理论及实践意义。为描述粗糙集，先介绍如下几个概念。

**定义 1**　设信息系统 IS=$(U,A,V,f)$，其中：$U=\{x_1,x_2,\cdots,x_n\}$为研究对象的非空有限集合，称为论域；$A=\{a_1,a_2,\cdots,a_n\}$是属性集合，$A=C\cup D$，$C$ 称为条件属性集，$D$ 称为决策属性集；$V=\bigcup_{a\in A}V_a$ 表示属性值的集合，$V_a$ 为属性 $a$ 的值域；$f:U\times A\to V$ 是一个信息映射函数，其为 $U$ 中每个对象 $x$ 赋予一个值，即 $\forall a\in A,x\in U$，有 $f_a(x)\in V_a$。若条件属性集 $C$ 与决策属性集 $D$ 满足：$A=C\cup D$ 且 $C\cap D=\phi$，则信息系统 IS 被称为决策表，而在属性约简后的决策表中，每一行记录均可抽象为一条规则。

**定义 2**　对任一属性集合 $R\subseteq C$，定义一个基于该属性的不可分辨关系，如式(2)所示。

$$\mathrm{IND}(R)=\{(x,y)\mid(x,y)\in U^2,\ \forall_{a\in R}(f(x,a)=f(y,a))\} \tag{2}$$

其中，IND($R$)是论域 $U$ 上的等价关系，而所有等价类集合记为 $U/\mathrm{IND}(R)$，等价类表示为：$[x]_{\mathrm{IND}(R)}=\{y\mid y\in U,(x,y)\in\mathrm{IND}(R)\}$

**定义 3**　对于信息系统 IS=$(U,A,V,f)$，令 $R\subseteq C$，$X\subseteq U$，则 $X$ 在等价关系 $R=\mathrm{IND}(R)$上的上近似集 $\bar{R}(X)$和下近似集 $\underline{R}(X)$定义分别如式(3)、式(4)所示。

$$\bar{R}(X)=\bigcup\{x\in U\mid[x]_R\cap X\neq\phi\} \tag{3}$$

$$\underline{R}(X)=\bigcup\{x\in U\mid[x]_R\subseteq X\} \tag{4}$$

其中，$[x]_R$ 是基于等价关系 $U/R$ 的等价类，即 $U/R=\{[x]_R\}=\{[x]_1,[x]_2,\cdots,[x]_n\}$。

**定义 4**　决策信息系统 DIS=$(U,A,V,f)$中，$U/R_D=\{D_1,D_2,\cdots,D_n\}$表示决策属性的划分，Dec($X$)为对象集合 $X$ 的描述，则决策信息系统导出的确定性决策规则 $D_i$ 如式(5)所示。

$$\varphi(\mathrm{Dec}(\underline{R_C}D_i))\to\varphi(\mathrm{Dec}(D_i)) \tag{5}$$

可以看出，决策类下近似集中描述的规则即为该决策的确定性规则。

## 2　面向领域文本的规则挖掘算法

基于 GMM 与粗糙集的文本规则知识发现过程：首先选取条件属性生成信息表；对于领域文本中存在的离散型与连续型并存的混合数据，本文通

过高斯混合模型聚类算法对连续型属性值进行离散化处理,同时也对信息表进行状态约简,并通过生成对应决策属性形成决策表;随后利用可辨识矩阵方法对决策表进行属性约简形成约简表;最后对约简表中潜在的知识规则进行挖掘和提取,导出决策规则。

### 2.1 决策信息表

通过对裁判文书的分析,本文将目标数据的起始位置、词性以及停止符数目等可能会对预测造成影响的各种因素作为条件属性,令条件属性:{起始位置,终止位置,长度,比例,停止符数目,词性}分别为 $\{c_1,c_2,c_3,c_4,c_5,c_6\}$,即 $C=\{c_1,c_2,c_3,c_4,c_5,c_6\}$。为了得到上述属性的值,本文相对目标数据进行标注:$\{s,e\}$,其中,$s$ 标注于目标数据的首部,$e$ 标注于目标数据的尾部。通过算法1来获取上述条件属性信息。

**算法1:getattributeinfo**

```
输入:标注的文本字符串 S
输出:数据的特征信息
Begin
  list=[ ]
  For s in S
    c1,c2=index(s,e)
    length=getlength(c1,c2)
    string=getsubstrs(c1,c2)
    for i in string:
      if tag(i)==“。”:
        number+1
    substring=gethead(s,c1)
    character=getspeech(substring)
    list.append([c1,c2,length,number,character])
  list=getproportion(list)
End
```

算法1首先利用已有的标签得到起始位置($c_1$)与终止位置($c_2$);然后通过 $c_1$ 与 $c_2$ 得到数据长度,并通过遍历字符的方式累积得到停止符的数量;再利用jieba工具包得到目标数据的词性,最后在遍历所有文本字符串之后通过长度数值相比得到比例信息。信息表如表1所示。

在表1中:$\{e1,e2,\cdots,e27\}$ 代表研究对象,$\{c_1,c_2,c_3,c_4,c_5,c_6\}$ 为条件属性,表中每一行都包含了该对象在各个属性上的取值。其中,起始位置、停止符数目以及词性等属性的取值为有限个数,将每一种属性取值视为状态,则起始位置、停止符数目以及词性等属性值均为有限状态集。为方便之后的运算,本文采用离散数值集合$\{0,1,\cdots,n\}$对起始位置状态值进行相应替代,即{<0,34>,<1,35>,<2,36>,<3,37>}。终止位置、长度、比例等三个属性均为连续型属性,文献[13]应用k-means聚类算法对数据离散化,但该算法对噪声及离群值特别敏感,且对于非凸数据与不规则形状的聚类则无法通过该算法有效解决。为此本文采用高斯混合模型(Gaussian Mixture Model,GMM)聚类算法对数据进行离散化处理,高斯混合模型是一种基于概率密度函数的聚类方法,其将数据划分为若干份,每一份均用一个Gaussian分布进行拟合,最后将各Gaussian分布融合,因此GMM能够拟合任意分布的数据,且能够对数据中的噪声及离群点进行很好的处理。通过GMM将目标数据划分为多个簇,并提取每个簇的簇标签,通过使用该标签来代替簇内的所有数据实现离散化。三个连续型数据的聚类示意图如图1所示。

**表1 信息表**

| $U$ | $c_1$ | $c_2$ | $c_3$ | $c_4$ | $c_5$ | $c_6$ |
|---|---|---|---|---|---|---|
| $e1$ | 34 | 54 | 20 | 0.010157 | 1 | nr |
| $e2$ | 37 | 80 | 43 | 0.022373 | 1 | ns |
| $e3$ | 34 | 63 | 29 | 0.014293 | 1 | nt |
| $e4$ | 36 | 72 | 36 | 0.019956 | 1 | ns |
| $e5$ | 36 | 74 | 38 | 0.019895 | 1 | ns |
| $e6$ | 35 | 73 | 38 | 0.020708 | 0 | nr |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| $e$ 25 | 35 | 71 | 36 | 0.018386 | 1 | ns |
| $e$ 26 | 36 | 78 | 42 | 0.022364 | 1 | b |
| $e$ 27 | 35 | 72 | 37 | 0.018592 | 1 | nt |

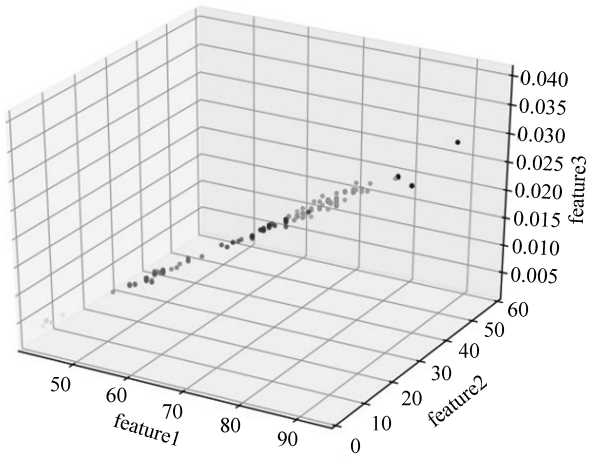


图1 连续型数据聚类图

图 1 是连续型数据的聚类示意图，坐标轴 feature1、feature2、feature3 分别代表条件属性 $c2$、$c3$、$c4$。从图中可以看出，利用 GMM 算法对连续型数据进行聚类划分，从左至右依次生成 5 个簇，分别用标签 $b_1$、$b_2$、$b_3$、$b_4$、$b_5$ 表示。令 $S=\{<b_i,d_i>|1\leqslant i\leqslant 5\}$为簇标签与数据的对应关系，$b_i$ 为簇标签，$d_i=\{d_{i1},d_{i2},\cdots,d_{in}\}$为簇 $b_i$ 内的数据。利用簇标签来替代簇内的所有数据，以此实现对连续数据的离散化处理。为构造决策属性，给出如下定义：

**定义 5**　在信息系统 IS 中，状态约简后的论域 $U=\{x_1,x_2,\cdots,x_i,\cdots,x_n\}$，对象集 $x_i=\{x_{i1},x_{i2},\cdots,x_{ij},\cdots,x_{in}\}$，其中 $x_i$ 为数据离散化后相同个体的合集，$x_{ij}$ 为对象集 $x_i$ 的一个源对象。令 $T(x_i)=\{T(x_{i1}),T(x_{i2}),\cdots,T(x_{in})\}$为对象集 $x_i$ 中所有源对象的 $c_2$ 属性集，同理 $L(x_i)$ 与 $P(x_i)$ 分别是对象集 $x_i$ 中所有源对象的 $c_3$ 与 $c_4$ 属性集。簇集合 clu=$\{T_m,L_m,P_m\}$表示条件属性 $c_2$，$c_3$ 和 $c_4$ 各簇领域的中间值，由程式化裁判文书启发性知识可以得出：

$$f(T)=\begin{cases}1,\mu(T(x_i))\leqslant \mathrm{tag}_i(T_m)\\0,\mu(T(x_i))>\mathrm{tag}_i(T_m)\end{cases}\tag{6}$$

$$f(L)=\begin{cases}1,\mu(T(x_i))\leqslant \mathrm{tag}_i(L_m)\\0,\mu(T(x_i))>\mathrm{tag}_i(L_m)\end{cases}\tag{7}$$

$$f(P)=\begin{cases}1,\mu(T(x_i))\leqslant \mathrm{tag}_i(P_m)\\0,\mu(T(x_i))>\mathrm{tag}_i(P_m)\end{cases}\tag{8}$$

$$\mathrm{decision}=\begin{cases}1,\mathrm{sum}(f(T),f(L),f(P))\geqslant 2\\0,\mathrm{sum}(f(T),f(L),f(P))<2\end{cases}\tag{9}$$

上式中，$\mu(T(x_i))$表示 $T(x_i)$的属性均值，$\mathrm{tag}_i(T_m)$代表对象集 $x_i$ 的属性 $c_2$ 所处簇的领域中间值，$\mathrm{sum}(f(T),f(L),f(P))$是求和函数。由领域启发式知识可以得到：当 $\mu(T(x_i))\leqslant\mathrm{tag}_i(T_m)$时，其效能（信息可靠性）为 1；当 $\mu(T(x_i))>\mathrm{tag}_i(T_m)$时，其效能为 0，同理 $f(L)$与 $f(P)$类似。由裁判文书的启发性知识可以得出，当 $\mathrm{sum}(f(T),f(L),f(P))\geqslant 2$ 时，该对象属性集效能高，代表其对数据描述的可靠性高，便于信息抽取，设为 1；反之，设置为 0。

根据定义 5 可以得到论域 $U$ 中各研究对象的决策属性，决策信息表如表 2 所示。

通过表 2 可以看出，利用 GMM 聚类算法对连续型数据进行离散化处理的过程中，由于同一簇内的数据均使用簇标签替代，使得不同的对象在同一属性上可能拥有相同的值（即簇标签），当若干个对象在所有属性上的值均相同时，则对象可合并为一个。这种方式能够缩减信息表的规模，即对其进行“状态约简”。其约简效果如表 3 所示。

**表 2　决策信息表**

| $U$ | 条件属性 | | | | | | 决策属性 |
|---|---|---|---|---|---|---|---|
| | $c_1$ | $c_2$ | $c_3$ | $c_4$ | $c_5$ | $c_6$ | $d$ |
| $x_1$ | 0 | $T_2$ | $L_2$ | $P_2$ | 1 | nr | 1 |
| $x_2$ | 3 | $T_4$ | $L_4$ | $P_4$ | 1 | ns | 0 |
| $x_3$ | 0 | $T_3$ | $L_3$ | $P_2$ | 1 | nt | 1 |
| $x_4$ | 2 | $T_3$ | $L_4$ | $P_4$ | 1 | ns | 1 |
| $x_5$ | 2 | $T_4$ | $L_4$ | $P_4$ | 1 | ns | 0 |
| $x_6$ | 1 | $T_3$ | $L_4$ | $P_4$ | 0 | nr | 1 |
| $x_7$ | 1 | $T_3$ | $L_4$ | $P_4$ | 1 | ns | 1 |
| $x_8$ | 2 | $T_4$ | $L_4$ | $P_4$ | 1 | b | 0 |
| $x_9$ | 1 | $T_3$ | $L_4$ | $P_4$ | 1 | nt | 1 |

**表 3　数据约简效果**

| 原始数据数目 | 约简后数据数目 | 约简效果/% |
|---|---|---|
| 27 | 9 | 66.7 |

## 2.2　基于可辨识矩阵的属性约简

属性约简是粗糙集中最重要的应用，也是决策规则提取最重要的手段。其目的是在保持决策表分类能力不变的情况下，剔除冗余属性，简化决策信息表，提高规则抽取的性能。本文采用可辨识矩阵的方法对决策表进行属性约简，其核心思想是保持可辨识矩阵的辨识能力不变，通过寻找全体条件属性集的一个子集，使该子集对决策属性的分类能力与全体条件属性集对决策属性的分类能力相同。定义如下：

**定义 6**　在决策信息系统 DIS=$\{U,A,V,f\}$，论域 $U=\{x_1,x_2,\cdots,x_n\}$为对象集合；$A=C\cup D$，其中 $C$ 和 $D$ 分别表示条件属性与决策属性；设论域 $U$ 依决策属性 $D$ 被划分为不同类簇，即 $\tilde{D}=\{X_1,X_2,\cdots,X_m\}$，则 DIS 中可辨识矩阵 $M(C)=\{m_{i,j}\}_{n\times n}$定义如式(10)所示。

$$m_{i,j}=\begin{cases}\phi, & x_i,x_j\in\tilde{D}\text{ 的同一价类}\\\{c\in C:f(c,x_i)=f(c,x_j)\}, & x_i,x_j\in\tilde{D}\text{ 的不同价类}\end{cases}$$
$$\text{其中，}1\leqslant i<j\leqslant n\tag{10}$$

基于可辨识矩阵的属性约简算法描述如下：

**Step1** 计算决策信息表 DIT 的可辨识矩阵 M(DIT)；

**Step2** 寻找决策信息表的核属性集，记为 $C_0$。将集合 $C_0$ 转为合取范式 $L=\bigwedge\limits_{C_k\in C_0} c_k$，其中 $c_k$ 为决策信息表中属性的核，$k=1,2,\cdots,n$，$n$ 是核属性个数。

**Step3** 在可识别矩阵中寻找不与集合 $C_0$ 拥有相同条件属性的元素集，表示为 $E_i$，即 $E_i\cap C_0=\varnothing$。其中，$i=1,2,\cdots,r$，$r$ 表示可辨识矩阵中与集合 $C_0$ 交集为空的元素个数。如果 $E_i$ 存在，则构建 $E_i$ 的析取表达式 $L_i=\bigvee\limits_{a_k\in E_i} a_k$。$a_k$ 元素为集合 $E_i$ 中的条件属性，$k=1,2,\cdots,m$，$m$ 为集合 $E_i$ 中的条件属性个数。对 $L_i(i=1,2,\cdots,r)$ 进行合取运算，得合取范式 $P=\bigwedge\limits_{E_i} L_i$。之后将合取范式 $P=\bigwedge\limits_{E_i} L_i$ 转为析取范式 $P=\bigvee L_i$，若 $E_i$ 不存在，则取 $P=1$。

**Step4** 对 $P$ 与 $L$ 进行合取运算 $P'=P\wedge L$。

**Step5** 将 $P'$ 转为析取范式 $P'=\bigvee Q$，$P'$ 可表示为：$P'=Q_1\vee Q_2\vee\cdots\vee Q_m$。

表 2 为决策信息表，依据上述可辨识矩阵的属性约简算法首先计算出决策信息表的可识别矩阵。由定义可知，可辨识矩阵是关于主对角线的对称矩阵，因此属性约简时可只考虑矩阵的上三角形式，如图 2 所示。

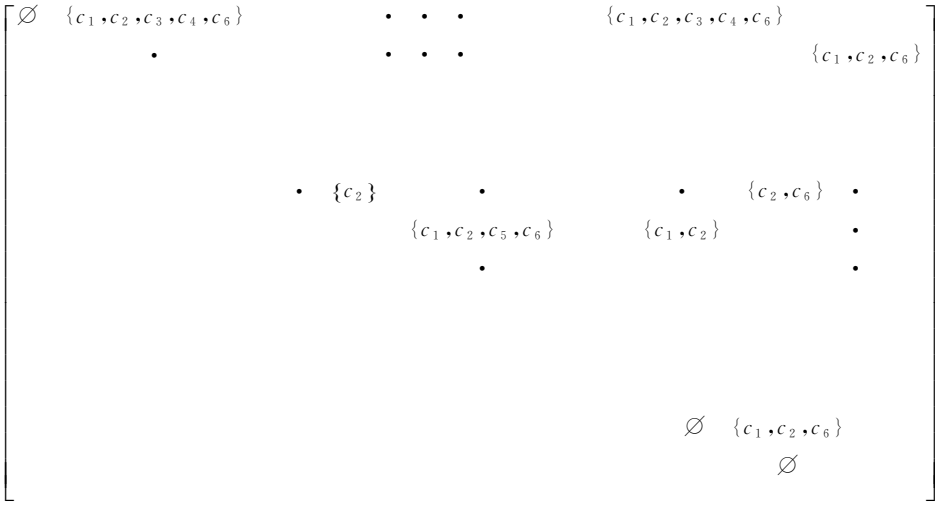


图 2 上三角矩阵图

可以由上式得出核属性集为 $C_0=\{c_2\}$，其合取范式即为其本身 $L=\{c_2\}$，又由于上式中的所有元素均与 $C_0$ 有交集，故 $P=1$。对 $P$ 与 $L$ 先进行合取运算，再进行析取运算，最后得析取范式 $P'=Q_1=\{c_2\}$，由此最后可得约简后的属性集合为 $\{c_2\}$。删除表 2 中的冗余属性后得最终决策表，如表 4 所示。

**表 4 约简决策信息表**

| $U$ | $c_2$ | $d$ |
|---|---|---|
| $x_1$ | $T_2$ | 1 |
| $x_2$ | $T_4$ | 0 |
| $x_3$ | $T_3$ | 1 |
| $x_4$ | $T_3$ | 1 |
| $x_5$ | $T_4$ | 0 |
| $x_6$ | $T_3$ | 1 |
| $x_7$ | $T_3$ | 1 |
| $x_8$ | $T_4$ | 0 |
| $x_9$ | $T_3$ | 1 |

### 2.3 文本规则的形成

在对决策信息表进行属性约简后，根据约简后的决策表，即可归纳导出决策规则。在上述示例中可得规则：$(C_2,T_2)\vee(C_2,T_3)\rightarrow(d,1)$。对决策表总结可以发现：条件属性 $c_2$ 对于决策属性有着决定性的影响，当 $c_2$ 属性为 $T_2$ 或 $T_3$ 值时，其效能为 1；当 $c_2$ 属性为 $T_4$ 值时，其效能为 0。结合先验知识进一步分析得出：司法领域的裁判文书是一种领域性很强的半结构化文本，其包含程式化的法言法语，对于文书的描述有较强的约束性，而通过上述属性约简可以发现裁判文书对于各类描述实体(即目标数据)的书写格式具有很强的规范性。

## 3 实验结果与分析

### 3.1 实验环境与评价指标

为了验证算法的有效性，本文使用司法领域的

裁判文书作为实验数据，分别利用本文算法得到的规则与传统方法对文本中的目标数据进行抽取，通过数据抽取的效果以及两种方法的对比来对本文算法进行评价。本实验平台将采用 Windows 10 64 位操作系统、CPU 为 Intel(R) Core(TM) i7-7500U，内存 8 GB 以及 PyCharm 集成开发环境。实验数据来自于中国裁判文书网，通过网络爬虫的方式获得。为比较本文算法与传统方法的优劣，本文将采用准确率 $P$、召回率 $R$ 以及 $F_1$ 值等指标来衡量信息抽取的效果。其定义如下：

(1) 准确率：正确预测为正在所有预测为正中所占的比例，如式(11)所示。

$$P=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \tag{11}$$

其中，TP 表示正例被判定为正例，也即抽取的信息为正确的数量；FP 表示负例被判定为正例，即抽取的信息为错误的数量。

(2) 召回率：正确预测为正占实际为正的比例，如式(12)所示。

$$R=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{12}$$

其中，FN 为正例被判定为负例，也即未被找到的正确信息的数量。

(3) $F_1$ 值：是准确率与召回率的加权平均，如式(13)所示。

$$F_1=\frac{2\times P\times R}{P+R} \tag{13}$$

### 3.2　实验结果

本次实验的数据来源于中国裁判文书网，共计 500 份裁判文书。分别用本文方法(GM-RS)与传统方法(TR-RU)对未标记的原始裁判文书进行抽取。其中，传统方法(TR-RU)主要是通过人工手段总结目标数据的结构“规律”，并提取规则，之后再利用这些规则进行数据抽取。在本次实验中，裁判文书抽取的内容(即目标数据)包括：原告、案件由来、判决正文以及受理费用。其中，原告(也称上诉人)主要描述案件发起者的姓名、性别以及民族等基本信息；案件由来主要描述案件名称与立案经过等内容；判决正文主要是法院法官依据双方当事人的辩诉以及证据对于案件做出判决；受理费用主要描述案件受理费的承担情况。抽取结果如表 5、表 6 所示。

从表 5、表 6 可以发现：相比于传统方法(TR-RU)，利用本文算法(GM-RS)得到的规则具有更高的抽取精度，其平均准确率以及 $F_1$ 值能够达到 95.0%和 95.7%。这是由于传统方法的抽取规则是人为手工定义，而裁判文书是一种复杂的半结构化领域文本，人为定义的规则在面对存在着动态性以及不确定性的领域文本时，无法对所有文本进行有效的匹配与信息抽取，同时传统方法还存在耗时、耗力以及需要研究人员具有一定专业知识背景等缺点。而本文方法能够在嘈杂的文本中提取出目标数据的关键属性，剔除噪声的干扰，通过属性约简后的决策表得到隐含的决策规则，利用这些规则进行高效的信息抽取，同时本文方法在耗费的时间与精力上也少于传统方法。两种算法数据抽取的对比图如图 3 所示。

**表 5　GM-RS 数据抽取结果表(%)**

| 目标数据 | 准确率 | 召回率 | $F_1$ 值 |
|---|---|---|---|
| 原告 | 88.75 | 90.0 | 89.37 |
| 案件由来 | 98.75 | 99.375 | 99.06 |
| 判决正文 | 95.0 | 96.25 | 95.62 |
| 受理费用 | 97.5 | 100 | 98.73 |

**表 6　TR-RU 数据抽取结果表(%)**

| 目标数据 | 准确率 | 召回率 | $F_1$ 值 |
|---|---|---|---|
| 原告 | 81.25 | 84.375 | 82.78 |
| 案件由来 | 85.625 | 86.875 | 86.25 |
| 判决正文 | 79.375 | 81.875 | 80.61 |
| 受理费用 | 91.875 | 93.75 | 92.80 |

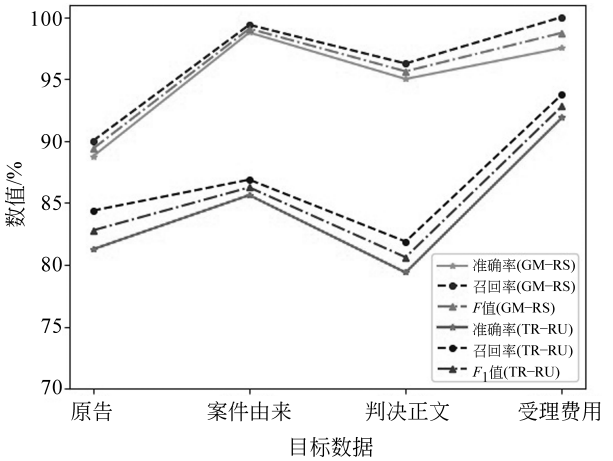


图 3　数据抽取对比图

从图 3 可以看出，本文算法在准确率、召回率以

及 $F_1$ 值方面的数值均优于传统方法。进一步研究可以发现：在两种方法中，原告以及判决正文的抽取精度均小于案件由来与受理费用。这是由于在裁判文书中，案件由来与受理费用属于“强结构”数据，其书写方式有比较严格的约束与限制。而相比之下，原告与判决正文则较为随意，这就导致了原告与判决正文这两类数据的“不确定性”增加，信息抽取的难度加大，最终导致抽取精度较低。

## 4 结束语

本文针对领域文本中的规则挖掘问题，提出了一种基于 GMM 与粗糙集的规则挖掘方法。对于领域文本中存在的混合型数据，该方法利用 GMM 聚类算法对连续型数据离散化处理，同时也对信息表进行纵向状态约简，并生成决策表；之后利用基于可辨识矩阵的属性约简算法对决策表进行属性约简得到最小属性约简，有效的减少了属性个数，简化了决策表结构；最后对约简后的决策表进行规则挖掘，得到规则集。在司法领域裁判文书上的实验结果表明：相比于传统的方法，决策规则对于裁判文书能进行很好的信息抽取，其平均准确率以及 $F_1$ 值能够达到 95.0%与 95.7%，同时决策规则的规模也较小。因此可以看出，本文提出的规则挖掘方法具有较大的实际意义。

## 参考文献


[1] Pawlak Z. Rough sets[J]. International Journal of Computer & Information Sciences, 1982, 11(5): 341-356.

[2] Chen L F, Tsai C T. Data mining framework based on rough set theory to improve location selection decisions: A case study of a restaurant chain[J]. Tourism Management, 2016, 53: 197-206.

[3] Shen K Y, Tzeng G H. Contextual improvement planning by fuzzy-rough machine learning: A novel bipolar approach for business analytics[J]. International Journal of Fuzzy Systems, 2016, 18(6): 940-955.

[4] Shao Y E, Chiu C C. Applying emerging soft computing approaches to control chart pattern recognition for an SPC-EPC process[J]. Neurocomputing, 2016, 201: 19-28.

[5] 李俊，孟涛，张立新，等. 基于粗糙集规则提取的导弹武器质量性能评估方法研究[J]. 兵工学报，2013，34(12)：1529-1535.

[6] 张腾飞，魏立力. 集中有序集值信息系统[J]. 计算机工程与应用，2014，50(16)：140-145.

[7] 卢娇丽，郑家恒. 基于粗糙集的文本分类方法研究[J]. 中文信息学报，2005，19(2)：67-71.

[8] Bezdek James C. Pattern recognition with fuzzy objective function algorithms[J]. Advanced Applications in Pattern Recognition, 1981, 22(1171): 203-239.

[9] Agrawal R, Gehrke J E, Gunopulos D, et al. Automatic subspace clustering of high dimensional data for data mining applications[J]. Data Mining and Knowledge Discovery, 1998, 27(2): 94-105.

[10] Ester M, Kriegel H P, Sander J, et al. A density-based algorithm for discovering clusters in large spatial databases with noise[C]//Proceeding of International Conference on Knowledge Discovery & Data Mining, 1996.

[11] Guha S, Rajeev Rastogi, Kyuseok Shim. Cure: An efficient clustering algorithm for large databases[J]. Information Systems, 1998, 26(1): 35-58.

[12] Theodoridis S, Koutroumbas K. Pattern Recognition, Fourth Edition[M]. Academic Press, 2008.

[13] 陈迎春，李鸥，孙昱. 基于聚类离散化和变精度邻域熵的属性约简[J]. 控制与决策，2018，33(08)：66-73.



洪壮壮(1994—)，硕士研究生，主要研究领域为机器学习、数据挖掘。
E-mail：1664260687@qq.com



万仲保(1965—)，硕士，副教授，硕士生导师，主要研究领域为知识工程、数据挖掘。
E-mail：zbwan@ecjtu.edu.cn



黄兆华(1966—)，通信作者，硕士，教授，硕士生导师，主要研究领域为知识工程、数据挖掘。
E-mail：hzh_nc@163.com