doi: 10.7690/bgzdh.2024.12.015

基于数据挖掘算法的犯罪相同特征向量集仿真

李凌君
（陕西警官职业学院信息技术系，西安 710021）

摘要：为提高犯罪数据挖掘准确率，有效打击犯罪，提出基于相同犯罪特征的数据挖掘算法。以公安系统数据库为基础，通过数据预处理，获得具有相同犯罪特征的数据向量集进行信息聚类，通过模糊集得到近似关系，依据贝叶斯公式将模糊集变为确定目标函数，依据目标函数，进行数据集划分相差较小的群组，获得数据挖掘结果。实验结果表明：该算法在犯罪特征的数据挖掘上，准确度和可靠性得到极大提升。

关键词：相同犯罪特征；数据挖掘；算法可靠性
中图分类号：TP274　　**文献标志码**：A

Simulation of Criminal Same Feature Vector Set Based on Data Mining Algorithms

Li Lingjun
(*Department of Information Technology*, *Shaanxi Police Vocational College*, *Xi'an* 710021, *China*)

Abstract: In order to improve the accuracy of crime data mining and effectively combat crime, a data mining algorithm based on the same crime characteristics is proposed. Based on the public security system database, through data preprocessing, a set of data vectors with the same criminal characteristics are obtained for information clustering. The approximate relationship is obtained through the fuzzy set, and the fuzzy set is transformed into a determined objective function according to the Bayesian formula. Based on the objective function, the data set is divided into groups with smaller differences to obtain data mining results. The experimental results show that the algorithm has greatly improved accuracy and reliability in data mining of criminal features.

Keywords: same criminal characteristics; data mining; algorithm reliability

0　引言

打击犯罪，保护人民生命财产安全，维护社会安定任务日渐繁重。随着互联网等信息技术的发展，我国公安行业积累了大量的数据，现已成为侦破犯罪、维护稳定的关键。如何采集到各类数据并进行快速准确分析，已成为公安信息系统领域研究的难题。数据挖掘就是从海量数据中快速挖掘信息，已成为目前研究热点[1-2]。

目前常用的数据挖掘算法[3-6]有聚类挖掘、关联规则挖掘等。其中聚类挖掘将研究对象以簇为单位进行划分，且簇内数据相似性高；关联规则挖掘主要用于挖掘数据集中的频繁模式。但大多数犯罪活动信息系统中，犯罪活动之间关系错综复杂，难以用简单的分组或关联规则进行表述，且数据挖掘准确率低。为提高数据挖掘信息的准确度，提高执法准确性，笔者根据原有公安系统数据中的相同犯罪特征数据，设计一种新型的数据挖掘算法。

1　基于相同犯罪特征的数据挖掘算法

笔者提供的数据挖掘算法以相同犯罪特征为基础，如图 1 所示。

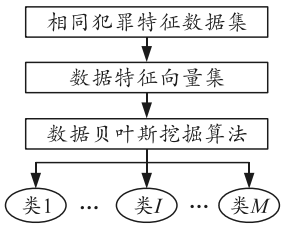

图 1　相同犯罪特征数据挖掘过程

由上图可将具体步骤包括：

1) 相同犯罪特征的数据聚类处理：具体以公安系统数据库为基础，通过数据预处理，获得具有相同犯罪特征的数据向量集进行信息聚类，其中数据为公安信息系统数据库 2000 年以来所有违法犯罪行为所具有的特征，包括性别、年龄、籍贯、出生年月、身高、专长、文化程度、职业等；

2) 构建目标函数：聚类后的信息通过模糊集得

收稿日期：2024-06-27；**修回日期**：2024-07-20
基金项目：陕西省教育厅专项科学研究计划资助(18JK0961)；陕西省教育厅一般专项科学研究项目“警务技战术实战化训练研究”(20JK0066)
第一作者：李凌君(1982—)，女，陕西人，硕士。

到近似关系，依据数据贝叶斯公式的后验概率，将模糊集变为确定目标函数；

3) 群组划分：依据目标函数，进行数据集划分相差较小的群组，获得数据挖掘结果。

1.1 相同犯罪特征数据聚类处理方法介绍

聚类分析是根据“物以类聚”道理进行分类的一种多元统计分析方法。是通过数学方法定量确定研究对象亲属关系及相似性而进行的划分。聚类分析的详细实现步骤如下：

1) 预处理，包括读取文本、分词、去除停用词等，笔者借助 NLPIR 汉语分词系统[7]实现。

2) 数据向量化。为进行相同犯罪特征数据向量化表示，笔者引入了 VSM 模型[8-9]，该模型以文本特征为维度属性，通过特征向量表示文本。

3) 规定特征权重 $q(t_a, d_b)$描述样本之间相似性，其中 $1 \geqslant q(t_a, d_b) \geqslant 0$，$q(t_a, d_b)=0$，表示 2 样本差异性大，$q(t_a, d_b)=1$ 表示 2 样本相似性高。

对于一个犯罪特征数据 d_x 来说，可表示成：

$$d_x=(t_{x1}, q_{x1}; t_{x2}, q_{x2}; \cdots; t_{xn}, q_{xn})。\quad (1)$$

式中：n 为特征项个数；t 为特征词；q 为特征词响应权重。具体向量构造步骤如下所示：

1) 将所有相同犯罪特征的数据用集合 D 表示，其中 $D=(d_1, d_2, \cdots, d_m)$，其中 d_m 表示某一相同犯罪特征的数据集；

2) 获得相同犯罪特征数据集中的特征词，并用集合 T 表示，$T=(t_1, t_2, \cdots, t_m)$；

3) 计算特征词权重 $q(t_a, d_b)$。关于特征词的权重计算公式，遵循式(2)[10]：

$$q(t_a, d_b)=\frac{tf(t_a, d_b)\cdot\log(N/n_i+0.01)}{\sqrt{\sum_{t_j\in d_b}(tf(t_a, d_b)\cdot\log(N/n_i+0.01))}}。\quad (2)$$

式中：$q(t_a, d_b)$为特征词 t_a 在相同犯罪数据特征 d_b 中的权重；$tf\ (t_a, d_b)$为特征词 t_a 在相同犯罪数据特征 d_b 中出现频率；N 为聚类数据总数；n_i 为数据集中出现特征词的数据总数。

1.2 构建目标函数

相同犯罪特征的数据具有相似关系，这种相似关系可以用函数表示，当相同犯罪特征的数据经过预处理后，利用模糊集思想[11-12]，通过 2 个近似集，包括上近似集和下近似集进行集合描述。同时，上下近似集之间有一个集合边界集，最外部边界以内为上近似集，外部表示外近似集。根据贝叶斯公式，将总体信息、先验信息、样本信息转换形成后验概率，根据后验概率进行数据分类。其中先验概率是指根据以后经验获得已知状态下的相同犯罪数据特征分布。依照式(3)构建目标函数[13]：

$$F=\sum_{n=1}^{N}\sum_{i=1}^{M}p_{ni}^{a}d^2(x_n, v_i)。\quad (3)$$

式中：N 为相同犯罪特征数总数；M 为聚类簇数目；a 为模糊因子，大于 1；v_i 为 i 类指标的中间距离；p_{ni} 为数据对象 x_n 隶属于第 i 类的后验概率；$d^2(x_n, v_i)$ 为欧几里得距离，$d^2(x_n, v_i)=\|x_n-v_i\|^2$。

1.3 数据挖掘

数据挖掘是指从海量数据中挖掘出隐含的、未知的、对决策者有潜在价值的关系、模式和趋势，笔者是以贝叶斯数据挖掘算法为基础进行分析，具体数据挖掘步骤如图 2 所示。

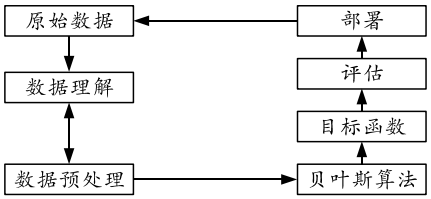

图 2　数据挖掘算法

对于从公安系统数据库中获得的相同犯罪特征数据经过预处理后，得到数据向量集，通过贝叶斯数据挖掘方法，将这些向量模糊聚类思想转换成目标函数，获得所需要挖掘的信息。贝叶斯数据挖掘算法是以总体信息、先验信息、样本信息为基础，通过贝叶斯公式转换形成后验概率，根据后验概率进行数据分类。其中先验概率是指根据以后经验获得已知状态下的相同犯罪数据特征分布。

2 仿真实验

为验证笔者设计的基于相同犯罪特征的数据挖掘算法的有效性，对某派出所提供的犯罪行为样本进行分析，具体样本数据如表 1 所示。

由表 1 可知，犯罪行为可以从犯罪程度较重和较轻 2 方面进行分析。本实验中用 C_1 表示犯罪程度较轻行为，C_2 表示犯罪行为较重行为。选取上表中的 10 个数据作为训练数据，如表 2 所示，其他数据作为测试数据，对表 2 进行训练。

用 P 表示事件出现的概率，对表 2 中的训练数据进行统计，确定 $P(C_1)=0.6$，$P(C_2)=0.4$；

对样本 X 依照相同犯罪特征(年龄=20～30，经济状况=中等，文化程度=初中，正当职业=无，犯罪记录=无，常住人口=是)进行分类，其中①P(年龄=20～30/C_1)=0.5，P(年龄=20～30/C_2)=0.74；②

P(经济状况=中等/C_1)=0.33，P(经济状况=中等/C_2)=0.25；③P(文化程度=初中/C_1)=0.5，P(文化程度=初中/C_2)=0.25；④P(正当职业=无/C_1)=0.333，P(正当职业=无/C_2)=0.5；⑤P(犯罪记录=无/C_1)=0.667，P(犯罪记录=无/C_2)=0.35；⑥P(常住人口=是/C_1)=0.5，P(常住人口=是/C_2)=0.5。

根据贝叶斯公式[14]：

$$p(C_1 / X) = p(X / C_1)p(C_1) \Big/ \sum_{j=1}^{n} p(X / C_1)p(C_1) 。 \quad (4)$$

计算得出各条件概率如表 3 所示。

表 1 某公安犯罪行为样本数据

年龄	经济状况	文化程度	正当职业	犯罪记录	常住人口	犯罪程度
20～30	中等	初中	无	无	是	较轻
＞40	差	小学	无	有	是	较重
30～40	差	初中	无	有	是	较重
20～30	中等	高中	有	无	是	较轻
＞40	差	小学	无	无	否	较重
30～40	差	初中	有	无	是	较轻
20～30	中等	初中	无	有	否	较轻
20～30	差	高中	无	无	否	较重
30～40	中等	高中	有	无	是	较轻
20～30	中等	初中	有	有	是	较重
20～30	差	高中	无	有	否	较重
＞40	差	初中	无	无	是	较轻
20～30	差	高中	有	无	否	较轻
20～30	差	高中	有	无	否	较轻
20～30	中等	高中	有	有	是	较轻

表 2 算法训练数据分析

年龄	经济状况	文化程度	正当职业	犯罪记录	常住人口	犯罪程度
＞40	差	小学	无	无	否	较重
30～40	差	初中	有	无	是	较轻
20～30	中等	初中	无	有	否	较轻
20～30	差	高中	无	无	否	较重
30～40	中等	高中	有	无	是	较轻
20～30	中等	初中	有	有	是	较重
20～30	差	高中	无	有	否	较重
＞40	差	初中	无	无	是	较轻
20～30	差	高中	有	无	否	较轻
20～30	差	高中	有	无	否	较轻

表 3 相同犯罪特征下样本数据条件概率

犯罪程度	年龄			经济状况		文化程度		
	20～30	30～40	＞40	中等	差	小学	初中	高中
较轻	0.50	0.33	0.12	0.33	0.67	0.001	0.51	0.49
严重	0.74	0.01	0.25	0.25	0.75	0.25	0.24	0.50
犯罪程度	正当职业		犯罪记录		常住人口			
	有	无	有	无	有	无		
较轻	0.33	0.67	0.33	0.67	0.50	0.50		
严重	0.50	0.50	0.75	0.25	0.25	0.75		

所以，$p(C_1/X)$=0.009 926；$p(C_2/X)$=0.002 928，即 $p(C_1/X)>p(C_2/X)$，分类结果为 C_1，即犯罪程度较轻，和表 1 吻合。如图 3 所示，为基于 C#语言程序设计如图 3 所示的界面。

为测试该算法的性能，与传统聚类分析算法[15]进行对比，由某一市公安系统内提取 1 200 个犯罪案件作为测试案例，通过比较测试结果与实际结案结果来进行预测准确率、目标函数值以及迭代次数等方面评价算法性能。表 4 为犯罪预测准确率，图 4 为迭代次数变化曲线。

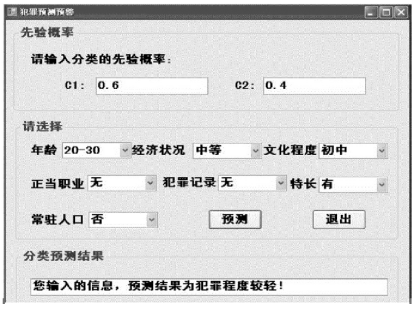

图 3 基于贝叶斯挖掘算法的犯罪预测

表 4　犯罪预测准确率

请补充表头	测试次数	预测成功数	准确率/%
改进的数据挖掘算法	1 200	1 101	91.7
传统聚类分析算法	1 200	940	78.3

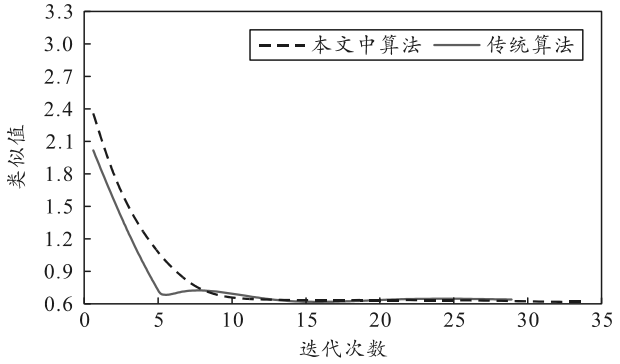

图 4　迭代次数

① 通过对该算法预测准确性进行测试，得到基于相同犯罪特征的数据挖掘算法预测准确性为 91.7%，而对于传统的聚类分析算法，笔者也对其预测准确性进行了测试，得到预测准确性为 78.3%。该算法的预测准确性高于传统聚类分析类数据挖掘算法。

② 本文中算法目标函数值计算结果和传统算法类似值均为 0.605 左右，但其迭代次数 34 次大于传统算法的迭代次数 28 次，说明本文中算法优于传统算法。

综上，验证了笔者设计的基于相同犯罪特征的数据挖掘算法在提高挖掘准确度和性能方面有一定提升。

3　结束语

公安机关现行的统计资料，难以对犯罪案件发生的轻重以及发案趋势和相关规律进行有效预测。由于公安行业的特殊性，特别是当前国际形势下，犯罪行为也出现了新的特点，如何挖掘有效信息进行预测成为关键。分类方法是数据挖掘中的重压方法之一，应用领域极为广泛。笔者基于相同犯罪特征，利用贝叶斯公式确定目标函数进行分类的数据挖掘算法预测准确性较高，能为公安机关侦破案件提供有力证据。这将会成为公安机关分析犯罪行为的重要研究方法。

参考文献：

[1] 冯姣. 大数据与犯罪侦查: 机遇, 挑战及应对[J]. 兰州学刊, 2019(5): 112-123.

[2] 唐昕森. 数据挖掘在计算机动态取证技术中的应用分析[J]. 信息通信, 2020(3): 160-161.

[3] 孙菲菲, 曹卓, 肖晓雷. 基于随机森林的分类器在犯罪预测中的应用研究[J]. 情报杂志, 2014, 33(10): 148-148.

[4] 侯德君. 对计算机数据挖掘技术的探讨[J]. 信息周刊, 2020(2): 1.

[5] 廖彬, 张陶, 于炯, 等. 多 MapReduce 作业协同下的大数据挖掘类算法资源效率优化[J]. 计算机应用研究, 2020, 037(5): 1321-1325.

[6] 马强. 基于关联规则的漏洞信息数据挖掘系统设计[J]. 现代电子技术, 2020, 43(5): 90-93, 99.

[7] 宫法明, 朱朋海. 基于自适应隐马尔可夫模型的石油领域文档分词[J]. 计算机科学, 2018, 45(S1): 110-113.

[8] 刘翔, 施千卫, 丁祖荣. 论文相似度的计算研究——基于 VSM 模型[J]. 情报杂志, 2010, 29(2): 142-144.

[9] 张传美. 基于聚类分析算法的舰船数据挖掘[J]. 舰船科学技术, 2020, 42(6): 170-172.

[10] 黄云, 洪佳明, 颜一鸣. 基于图的特征词权重算法及其在文档排序中的应用[J]. 计算机系统应用, 2012, 21(6): 216-219, 194.

[11] 吴德垠, 杨高进. 闭 G-V 模糊拟阵的模糊圈公理[J]. 吉林大学学报(理学版), 2020, 58(2): 239-250.

[12] 罗天正, 关皓. 政治关联, 营商环境与企业创新投入——基于模糊集定性比较分析[J]. 云南财经大学学报, 2020, 36(1): 69-79.

[13] 吴绍兵, 王昌梅. 一种基于贝叶斯和决策树的少数民族犯罪数据挖掘方法比较研究[J]. 计算机科学与应用, 2019, 9(2): 339-350.

[14] 李艳美, 张卓奎. 基于贝叶斯网络的数据挖掘方法[J]. 计算机仿真, 2008(2): 87-89.

[15] 王德青, 朱建平, 刘晓葳, 等. 函数型数据聚类分析研究综述与展望[J]. 数理统计与管理, 2018, 37(1): 55-67.