同源数据的协同挖掘算法研究

王 泳 吕 科 潘卫国
(中国科学院大学 北京 100049)

摘 要 围绕知识管理和提高数据挖掘模型的可解释性问题展开研究,提出了采用协同挖掘的方法对同源数据进行模式评估和知识管理的CMA算法(Collaborative Mining Algorithm)。与集成学习产生同一类型知识规则的组合学习方式不同,协同挖掘在同源数据的基础上建立不同类型的学习模型,并且每类学习模型产生的知识规则的表现形式各不相同,通过比对学习形成了一致的知识规则。实验表明,协同挖掘可以有效发现数据中的隐含信息,提高知识管理的性能。

关键词 同源数据,协同挖掘,模型评估,知识管理
中图法分类号 TP182 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2014.12.030

Research on Collaborative Mining Algorithm on Homologous Data

WANG Yong LV Ke PAN Wei-guo
(University of Chinese Academy of Sciences, Beijing 100049, China)

Abstract This article explored the issues of knowledge management and improvement of the interpretability of data mining models, and proposed the collaborative mining algorithm (CMA), which performs pattern evaluation and knowledge management based on collaborative mining of homologous data. In contrast to the ensemble learning knowledge rules by combining learning models of the same type, collaborative mining sets up learning models of different types based on homologous data, and each model owns different forms of knowledge rules. Through the comparison study, coincident knowledge rules were formed. Experiments show that collaborative mining can efficiently find the latent information in data, and improve the performance of knowledge management.

Keywords Homologous data, Collaborative mining, Model evaluation, Knowledge management

1 引言

作为知识发现(Knowledge Discovery in Database, KDD)的一个重要环节[1],数据挖掘(Data Mining, DM)是从大量数据中挖掘出有趣模式和知识的过程[2]。传统的数据分析方法(例如统计方法)只能获得数据的表层信息,而数据挖掘可以寻找和观察数据之间的某种规律,增强数据的"可解释性"[3]。自上世纪90年代这一学科建立起来,众多学者在这一领域开展了很多有意义的研究工作,使得数据挖掘算法模式日益增多,任务功能日趋繁杂,从数据中所能获取的知识表达形式也逐渐丰富起来[4]。

数据挖掘的基本任务功能大体可以分为两类[5],在每种功能下都可以衍生出数十种算法(如图1所示),如决策树、统计学习、贝叶斯算法、神经元网络等便是使用范围比较广的几类数据挖掘算法[6]。但随着数据挖掘模式和知识的增多,如何评估、管理这些知识模型已成为当前日益紧迫的研究课题[7]。

本文重点研究了对不同任务功能下不同算法模式发现的知识模型进行评估和管理的问题,第2节在文献调研的基础上提出了同源数据和协同挖掘的概念并进行了对比讨论;第3节提出了对同源数据采用协同挖掘的方法进行模式评估和知识管理的CMA(Collaborative Mining Algorithm)算法;第4节是数据实验;最后对文章内容进行了总结。

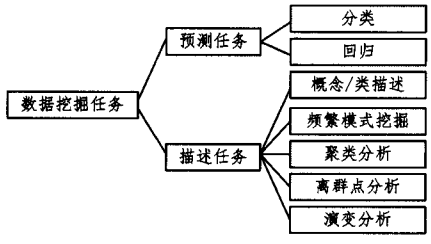


图1 数据挖掘的基本任务功能

2 模型评估与知识管理

2.1 同源数据

数据挖掘知识模型的建立一般需要经过以下几个主要步骤:

1)确定和逐步理解应用领域;

2)数据准备,对原始数据进行数据预处理;

到稿日期:2014-01-15 返修日期:2014-05-10 本文受国家自然科学基金(61371155)资助。

王 泳 博士后,讲师,主要研究方向为知识发现、数据挖掘、模式识别,E-mail:wangyong@ucas.ac.cn;吕 科 教授,博士生导师,主要研究方向为知识发现、数字图象处理;潘卫国 博士生,主要研究方向为知识发现、数字图象处理。

3)开发模型、构建假设；

4)选择合适的数据挖掘算法建立挖掘模型；

5)结果分析、知识模型的解释和评估；

6)知识管理。

原始数据一般是不能直接用来建立数据挖掘模型的，这不仅是因为原始数据集中可能存在大量数据噪声或数据缺失，更重要的是针对不同数据挖掘算法适用的数据类型是不一样的，所以在实施某一类数据挖掘算法前都需要进行数据预处理[8,9]。数据预处理在数据挖掘过程中起到基础性作用，一般会占整个工作量的60%。数据预处理通常包括数据清洗、数据集成、数据规约、数据变换、数据标准化、数据离散化等过程[10]。

定义1 原始数据集 D 经过数据预处理后形成新的数据集 D_1 和 D_2，若 $D_1 \neq D2$，则 $D1$ 与 $D2$ 互称为同源数据。

2.2 协同挖掘

根据“没有免费的午餐定理”(No Free Lunch，NFL)[11]，在没有任何先验知识的情况下，只根据训练数据所获得的学习模型是不可靠的，但结合先验知识采用数据变换后的训练数据所学习得到的模型，其可靠性完全可以得到保证。这主要是因为不同学习算法之间并没有天生的优劣之分，只是适用的假设条件和要求不同。同一类型学习算法的算法基础是一样的，但不同类型学习算法之间是不同的，尤其是担任不同任务功能的算法之间的差异性更大(如表1所列)。

表1 不同类型学习算法的差异对比

学习算法	算法基础	任务功能
决策树	信息论	分类、回归
贝叶斯	概率论	分类
神经元网络	仿生学	分类、回归
数据立方体模型	数据库	概念/类描述
Apriori算法	数据结构	频繁模式挖掘
k-均值算法	迭代重定位	聚类分析
基于似然的检测模型	统计学	离群点分析
隐马尔科夫模型	概率论	演变分析

通用的模型评估方法是在置信度、置信区间框架之下采用不同模型评价准则对同一数据在不同算法下的挖掘效果进行评估[12]，得出的知识模型也是在同一类任务功能下进行验证，缺乏其他知识模型的校正。实施不同任务功能的学习算法会从不同角度对同一数据中隐含的知识进行挖掘、验证和归纳。虽然知识规则的展示形式不一样，学习算法的效率有差异，但知识模型之间并没有优劣之分，只是视角不同，通过相互校正可以提炼知识，实现知识的有效管理。

定义2 对于同源数据 D_1 和 D_2 分别使用不同类型数据挖掘算法 A_1 与 A_2 建立不同的算法模型 M_{AD1} 与 M_{AD2}，如果根据算法模型得出的知识规则 LM_{AD1} 与 LM_{AD2} 是一致的，则 A_1 与 A_2 互称为协同挖掘算法，这一学习过程被称为协同挖掘。

2.3 协同挖掘与知识管理

知识管理(Knowledge Management，KM)是自上世纪90年代中期开始在全球崛起的一项学术与商业应用主题，它通过管理与技术手段，使人与知识紧密结合[13]。知识管理可以实现知识的有效利用，它针对个人及社群所拥有的显性知识和隐性知识的确认、创造、掌握、使用、分享及传播进行积极有效的管理。

知识的管理应用按照层级可以分为数据、信息、知识及智慧4个阶段(如图2所示)。数据挖掘是知识管理的技术基础，通过数据挖掘可以从数据中得到有用的隐性信息，但这些信息还不能称为知识，因此数据挖掘本身并不产生知识，也不能进行数据管理。与传统模式评估只能基于单一挖掘模型框架进行知识获取和管理不同，协同挖掘可以通过多个挖掘模型之间的比对学习，实现多类型知识的融合，因此协同挖掘是是实现知识管理的一种技术手段。

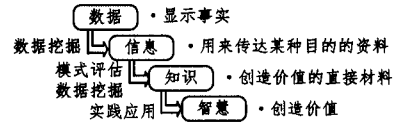


图2 知识应用的层次

2.4 协同挖掘与集成学习

集成学习(Ensemble Learning)是数据挖掘领域中一类应用广泛的算法模型，它通过集成一些弱学习算法形成一个统一的算法模型。基于弱学习理论，通过集成学习得到的算法模型与强学习算法模型等价，但模型泛化能力会更强[14]。集成学习的最简单方法是通过投票(voting)的方式线性组合基学习器。假设有 L 个基学习器，以 $d_j(x)$ 表示基学习器 M_j 在给定的任意输入 x 上的预测结果，w_j 表示基学习器 M_j 的权重，那么线性组合集成学习的总体输出 y 可以表示为：

$$y=\sum_{j=1}^{L} w_j d_j(x) \tag{1}$$

其中，$w_j \geqslant 0$ 且 $\sum_{j=1}^{L} w_j = 1$。

假设基学习器 $d_j(x)$ 之间相互独立同分布，且期望为 $E[d_j(x)]$，方差为 $Var[d_j(x)]$，那么当使用 $w_j = 1/L$ 时，输出的期望和方差为：

$$E[y]=E[\sum_{j=1}^{L}\frac{1}{L}d_j(x)]=\frac{1}{L}LE[d_j(x)]=E[d_j(x)] \tag{2}$$

$$Vary[y]=Var[\sum_{j=1}^{L}\frac{1}{L}d_j(x)]=\frac{1}{L^2}Var[\sum_{j=1}^{L}d_j(x)]$$

$$\frac{1}{L^2}LVar[d_j(x)]=\frac{1}{L}Var[d_j(x)] \tag{3}$$

从式(2)、式(3)看出，集成学习后的总体期望没有改变，因而偏差也未改变，但方差随着基学习器数量 L 的增加而降低(即学习结果更加稳定，模型泛化能力更强)[11]。在一般情况下，如果基学习器之间并非独立而是负相关，则从式(4)可以看出，集成学习后的总体方差可以进一步降低。如果随之增加的偏差不是很高，则误差也会降低。

$$Var(y)=\frac{1}{L^2}Var(\sum_{j=1}^{L}d_j(x))$$

$$=\frac{1}{L^2}(\sum_{j=1}^{L}Var(d_j(x))+2\sum_{j=1}^{L}\sum_{i<j}Cov(d_j(x),d_i(x))) \tag{4}$$

从集成学习算法模型中得出的知识规则是在一个集成了若干弱学习算法的统一的算法模型中产生的，这些知识规则无法从单一弱学习算法中得到，因此集成学习强调知识的发现。

协同挖掘是通过不同学习模型的比对学习形成一致的知识规则，每个学习模型都可以各自产生知识规则，因此协同挖

掘强调对知识的管理。

3 CMA 算法

协同挖掘并不只是建立数据挖掘的算法模型，而是更加强调对算法模型的知识管理，因此实现协同挖掘的 CMA 算法包括两个过程：1)建立知识规则，2)管理知识规则。具体算法如图 3 所示。

算法：CMA. 使用协同挖掘实现知识管理

输入：D：原始数据集；A_i：不同类型的学习算法；Max_L：最大知识规则数。

输出：LM：知识规则库

方法：

1. 按以下方法建立知识规则

1)对原始数据集 D 进行数据预处理，产生同源数据集 D_i

2)将学习算法 A_i 应用到同源数据集 D_i 建立算法模型 M_{AD_i}

3)从算法模型 M_{AD_i} 中建立知识规则集合 LM_{AD_i}

2. 按以下方法管理知识规则

1)$L_i=0$，知识规则库中规则计数清零

2)Do while $LM_{AD_i}\neq\emptyset$ and $L_i<$Max_L

3)对不同知识规则 LM_{AD_i} 进行知识比对学习

4)If LM_{AD_i} 比对成功

5) $L_i=L_i+1$

6) LM 中放入知识规则 LM_{AD_i}

7)End

8)集合 LM_{AD_i} 中删除已经比对学习过的知识规则

9)Loop

图 3 实现协同挖掘的 CMA 算法

CMA 算法的算法复杂性由两部分组成：在建立知识规则阶段，算法的复杂性是每个单一学习模型算法复杂性的线性组合，因为每个单一学习模型是基于同源数据建立模型，相互之间不存在数据纠缠和算法迭代，所以可以通过并行计算的方式降低算法复杂性；在管理知识规则阶段，算法的复杂性与知识规则库的容量和参与比对学习的规则数量相关，因为参与比对学习的规则是两两比对，所以比对的次数是规则数量的平方，这时可以通过添加知识规则库的容量限制，对算法复杂性加以约束。

4 数据实验

协同挖掘是通过不同学习模型的比对学习形成一致的知识规则，因此需要建立不同任务功能的学习模型，而担任不同任务功能的数据挖掘算法所适用的数据类型是不一样的，因此在进行协同挖掘之前都需要将数据转化为同源数据，而这种转化对大多数结构化数据都是可行的。为了说明算法的通用性，实验将在开源的数据集和实验环境上实施。

实验环境采用怀卡托智能分析环境(Waikato Environment for Knowledge Analysis，Weka)系统(http://www.cs.waikato.ac.nz/ml/weka)。

实验 1 实验数据来自 UCI 公开数据库中 iris(鸢尾花)数据集(http://archive.ics.uci.edu/ml/datasets/Iris)。数据集中有 150 个样本，平均分为 3 类(Iris-setosa，Iris-versicolor，Iris-virginica)，每个样本有 5 个属性特征，属性描述如表 2 所列。

表 2 iris(鸢尾花)数据集属性描述

属性	类型	最小值	最大值	均值	方差
sepallength (萼片长度)	数值型	4.3	7.9	5.843	0.828
sepalwidth (萼片宽度)	数值型	2	4.4	3.054	0.434
petallength (花瓣长度)	数值型	1	6.9	3.759	1.764
petalwidth (花瓣宽度)	数值型	0.1	2.5	1.199	0.763
class(类)	名词型	Iris-setosa，Iris-versicolor，Iris-virginica			

1) 使用 J48 决策树算法进行分类实验，挖掘知识规则。

从图 4(a)中可以推导出 5 条规则(正确分类 147 个样本，错分 3 个样本)：

If petalwidth $<=0.6$，then class is Iris-setosa

If $0.6<$ petalwidth $<=1.7$ and petallength $<=4.9$，then class is Iris-versicolor

If $0.6<$ petalwidth $<=1.5$ and $4.9<$ petallength，then class is Iris-virginica

If $1.5<$ petalwidth $<=1.7$ and $4.9<$ petallength，then class is Iris-versicolor

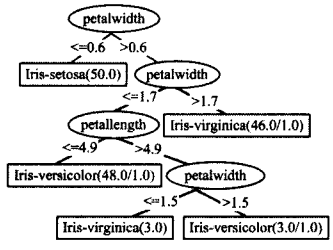
If $1.7<$ petalwidth，then class is Iris-virginica

从图 4(b)中可以推导出 3 条规则(正确分类 144 个样本，错分 6 个样本)：

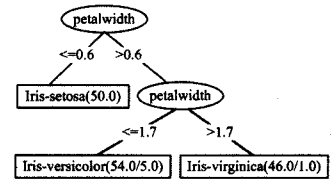
If petalwidth $<=0.6$，then class is Iris-setosa

If $0.6<$ petalwidth $<=1.7$，then class is Iris-versicolor

If $1.7<$ petalwidth，then class is Iris-virginica



(a) 分支最小样本数为 2



(b) 分支最小样本数为 10

图 4 使用 J48 算法分类 Iris 数据集产生的决策树

虽然图 4(b)中决策树的分类精度比图 4(a)略有下降，但由于使用了要求更高的剪枝约束(分支最小样本数)，因此形成的规则更容易理解，显示出属性“petalwidth”的重要性。

2)使用 Apriori 算法进行关联规则实验，挖掘知识规则。

由于关联规则挖掘无法处理连续属性，因此使用等值装箱方法将前 4 个数值属性分别离散成 10 个区间，离散结果如表 3 所列。

设置置信度为 1，最小支持度为 $0.006\approx1/150$(即至少要有 1 个样本满足该条规则)。如果全部属性都参与关联计算，会产生 2340 条规则，其中大多数都是冗余规则且处理复杂。注意到在以上决策树产生的规则中属性“petalwidth”和“pet-

allength"具有重要意义，因此单独考虑这两个属性与类标的关联关系。

考虑属性"petalwidth"与"class"的关联关系，产生 7 条规则，合并后得出 3 条规则：

petalwidth='(-inf-0.82]'==> class=Iris-setosa (50 个样本满足该规则，即 Iris-setosa 类被完全正确识别)

petalwidth='(0.82-1.3]'==> class=Iris-versicolor (28 个样本满足该规则)

petalwidth='(2.02-inf]'==> class=Iris-virginica (23 个样本满足该规则)

表 3 iris(鸢尾花)数据集数值属性离散化

sepallength (萼片长度)		sepalwidth (萼片宽度)		petallength (花瓣长度)		petalwidth (花瓣宽度)	
离散标签	样本数	离散标签	样本数	离散标签	样本数	离散标签	样本数
(-inf-4.66]	9	(-inf-2.24]	4	(-inf-1.59]	37	(-inf-0.34]	41
(4.66-5.02]	23	(2.24-2.48]	7	(1.59-2.18]	13	(0.34-0.58]	8
(5.02-5.38]	14	(2.48-2.72]	22	(2.18-2.77]	0	(0.58-0.82]	1
(5.38-5.74]	27	(2.72-2.96]	24	(2.77-3.36]	3	(0.82-1.06]	7
(5.74-6.1]	22	(2.96-3.2]	51	(3.36-3.95]	8	(1.06-1.3]	21
(6.1-6.46]	20	(3.2-3.44]	18	(3.95-4.54]	26	(1.3-1.54]	20
(6.46-6.82]	18	(3.44-3.68]	9	(4.54-5.13]	29	(1.54-1.78]	6
(6.82-7.18]	6	(3.68-3.92]	11	(5.13-5.72]	18	(1.78-2.02]	23
(7.18-7.54]	5	(3.92-4.16]	2	(5.72-6.31]	11	(2.02-2.26]	9
(7.54-inf)	6	(4.16-inf)	2	(6.31-inf)	5	(2.26-inf)	14

注：inf 表示无穷大

考虑属性"petallength"与"class"的关联关系，产生 7 条规则，合并后得出 3 条规则：

petallength='(-inf-2.18]'==> class=Iris-setosa (50 个样本满足该规则，即 Iris-setosa 类被完全正确识别)

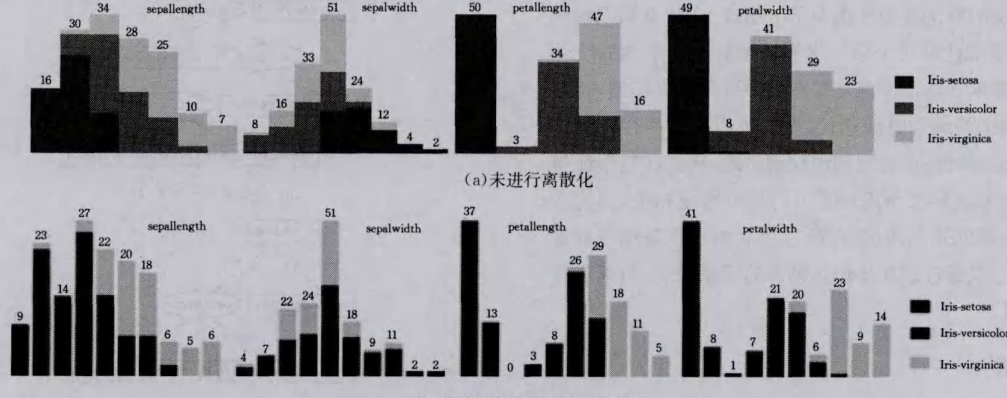
petallength='(2.77-3.95]'==> class=Iris-versicolor (11 个样本满足该规则)

petallength='(5.13-inf]'==> class=Iris-virginica (34 个样本满足该规则)

从以上 6 条规则可以看出，只要单独使用属性"petalwidth"或"petallength"就可以将"Iris-setosa"类完全正确识别，但将"Iris-versicolor"类和"Iris-virginica"类完全正确识别出来单独依靠属性"petalwidth"或"petallength"是不行的。

3) 使用可视化技术挖掘知识规则。

从决策树和关联规则挖掘得出的知识规则可以看出，"Iris-setosa"类比较容易识别，属性"petalwidth"或"petallength"在识别中起到更重要的作用，为了增加这些规则的"可解释性"和"透明度"，绘制属性与类别的分布关系图(如图 5 所示)。



(a)未进行离散化

(b)等值装箱离散成 10 个区间

图 5 Iris 数据集中属性与类别的分布关系图

图 5(a)表明 3 类样本在 4 个属性上的同值区间分别存在较多重叠区域，相对而言，属性"sepallength"和"sepalwidth"的重叠区域和覆盖样本较多，不便识别分类。图 5(b)表明经过数值离散化后这一情况并没有改变，但属性"petallength"和"petalwidth"的重叠区域和覆盖样本相对减少，分布也更加清晰。"Iris-setosa"类使用属性"petalwidth"或"petallength"就可以完全正确识别，"Iris-versicolor"类使用属性"petalwidth"可以正确识别大多数样本，"Iris-virginica"类使用属性"petallength"可以正确识别大多数样本，"Iris-versicolor"类和"Iris-virginica"类较难识别的样本主要集中分布在数值中间区域。

实验 2 本实验数据来自 CMU 公开数据库中 quake(地震)数据集(http://lib.stat.cmu.edu/datasets/smoothmeth)。数据集中有 2178 个样本，每个样本有 4 个属性特征，属性描述如表 4 所列。

表 4 quake(地震)数据集属性描述

属性	类型	最小值	最大值	均值	方差
focal_depth (震源深度)	整数型	0	656	74.36	116.47
latitude(纬度)	数值型	-66.49	78.15	7.96	30.55
longitude(经度)	数值型	-179.96	180	54.915	118.88
richter(里氏震级)	数值型	5.8	6.9	5.98	0.19

1)使用回归算法进行回归实验,挖掘知识规则。

为了消除计量单位对回归误差的影响,首先对所有数据进行标准化处理,然后使用线性和非线性等多种回归算法对地震数据进行规则挖掘(如表5所列)。

表5 quake(地震)数据集回归

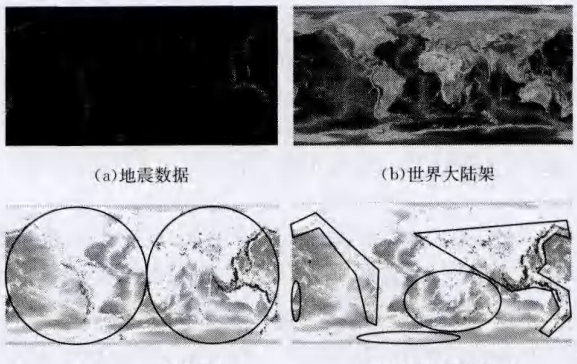
回归方法	相关系数	平均绝对误差	相对绝度误差
线性回归	0.0766	0.1484	99.49%
神经元网络(单隐层2神经元)	0.087	0.2091	140.18%
神经元网络(单隐层5神经元)	0.1393	0.2081	139.54%
SMO(多项式核)	0.0763	0.1407	94.34%
SMO(RBF核)	0.0679	0.1406	94.31%
M5P(回归决策树)	0.2011	0.1473	98.79%

以上回归模型显示出震级与震源深度、纬度、经度这3种属性的相关性很小,回归预测的精度很差。这主要是因为数据属性太少,根据现有数据所包含的信息量无法准确预测震级。

2)使用可视化技术挖掘知识规则。

从回归模型中可以得出地震发生的强度与地理位置没有太大关联的规则,但地震发生与地理位置是否有关联可以使用可视化工具进行挖掘。

图6(a)显示了地震发生的位置信息,将其与图6(b)世界大陆架的位置信息进行对比可以看出,地震发生的位置基本围绕在世界大陆架的周围。



(a)地震数据 (b)世界大陆架

(c)X-means聚类效果 (d)DBSCAN聚类效果

图6 地震数据集与世界大陆架地震带对比图

3)使用X-means[15]和DBSCAN算法进行聚类分析,挖掘知识规则。

使用可视化工具可以得出地震发生与地理位置有很大关联的规则,但地理位置之间是否存在关联规律可以使用聚类分析的方法进行挖掘。为了消除其他因素影响,数据预处理后只保留纬度(latitude)和经度(longitude)属性进行聚类分析。

X-means算法是对K-means算法进行改进的基于划分的聚类算法,它可以根据预先设定的簇的取值范围择优选择聚类结果。基于划分的聚类方法都是根据距离进行聚类的,因此聚类形成的簇通常是球形。本次实验预先设定簇的取值范围在2~8之间,实验最终的聚类结果是形成2个球形的簇,较好地吻合世界大陆架的东、西两个半球的分布,聚类效果如图6(c)所示。

DBSCAN算法是基于密度的聚类算法,聚类形成的簇通常没有固定的形状,只具有高密度的连通区域。本次实验预先设定密度圆的半径参数epsilon=0.1,圆内最小样本个数参数minPoints=20,实验最终的聚类结果是形成5个连通簇,另有36个样本没有形成簇,聚类效果如图6(d)所示。相比而言,使用DBSCAN算法形成的簇更加符合环太平洋美洲地震带、环太平洋亚欧地震带、印度洋地震区、南极地震带和澳洲东部地震带等地震学上已知的地震带理论。

结束语 本文从知识管理的角度出发,阐述了不同数据挖掘算法模型在同源数据的基础上开展协同挖掘的重要意义和实现方法。与传统的建立单一数据挖掘算法模型的研究方法不同,本文提出的CMA协同挖掘算法并不关注单一算法模型的精度能有多高,而是更加强调知识规则的比对学习,增强挖掘模型的"可解释性"、"透明度"和"可靠性"。通过协同挖掘可以增强模型的适用范围,这对理解不同挖掘模型本质的问题是有益的探索。未来的研究将更加关注算法模型之间的协同互补,以及知识规则库的可伸缩性。

参考文献

[1] Fayyad U M, Shapiro G P, Smyth P. The KDD Process for Extracting Useful Knowledge from Volumes of Data[J]. Communications of the ACM, 1996, 39(11): 27-34

[2] Han Jia-wei, Kamber M, Pei Jian. Data Mining: Concepts and Techniques(3rd edition)[M]. Singapore, Elsevier, 2012

[3] 郭萌,王珏. 数据挖掘与数据库知识发现:综述[J]. 模式识别与人工智能,1998,11(3):292-299

[4] 胡包钢,王泳,杨双红,等. 如何增加人工神经元网络的透明度?[J]. 模式识别与人工智能,2007,20(1):72-84

[5] Tan Pang-ning, Steinbach M, Kumar V. Introduction to Data Mining[M]. Addison Wesley, 2005

[6] Mitra S, Pal S K, Mitra P. Data Mining in Soft Computing Framework: A Survey[J]. IEEE Trans. on Neural Networks, 2002, 13(1): 3-14

[7] Lee M R, Chen T T. Revealing research themes and trends in knowledge management: From 1995 to 2010[J]. Knowledge-Based Systems, 2012, 28(4): 47-58

[8] West M. Developing High Quality Data Models[M]. Singapore, Elsevier, 2011

[9] 郭晓波,赵书良,刘军丹,等. 基于概念图的关联规则知识表示[J]. 计算机科学,2013,40(8):261-265

[10] 王泳,邢红杰. 对基于知识发现的神经元网络集成方法的研究[J]. 计算机科学,2006,33(10):189-192

[11] Duda R O, Hart P E, Stork D. Pattern Classification (2nd edition)[M]. New York, John Willy, 2001

[12] 王泳,胡包钢. 应用统计方法综合评估核函数分类能力的研究[J]. 计算机学报,2008,31(6):942-952

[13] Liberona D, Ruiz M, Fuenzalida D. Customer Knowledge Management in the Age of Social Networks[J]. Advances in Intelligent Systems and Computing, 2013, 172: 353-364

[14] 张春霞,张讲社. 选择性集成学习算法综述[J]. 计算机学报,2011,34(8):1399-1410

[15] Pelleg D, Moore A W. X-means: Extending K-means with Efficient Estimation of the Number of Clusters[C]// Seventeenth International Conference on Machine Learning. 2000: 727-734