

基于 Hilbert 空间理论的图像知识发现^{*})

游福成^{1,2} 杨炳儒²

(北京印刷学院信息与机电工程学院 北京 102600)¹ (北京科技大学信息工程学院 北京 100083)²

摘要 最近几年,知识发现研究的进展很快。目前,在知识发现领域图像数据知识发现形成了新的研究热点。本文介绍了基于 Hilbert 空间理论的图像知识发现模型 IMDFSSM,采用模式(定义为 Hilbert 空间中的矢量)来定量地表征图像数据的知识表示和参与知识发现过程。然后用图像挖掘系统作为实例进行了验证,结果表明该模型对于图像数据的知识发现过程具有指导性作用。

关键词 数据挖掘,图像挖掘,模式

Knowledge Discovery of Image Based on Hilbert Space Theory

YOU Fu-Cheng^{1,2} YANG Bing-Ru¹

(Institute of Graphic Communication Information & Machine-Electronic School of Beijing, Beijing 102600)¹

(School of Information Engineering, University of Science & Technology Beijing, Beijing 100083)²

Abstract The research of Knowledge Discovery has developed rapidly in recent year. Knowledge discovery of image data has become the research hotspot in KDD field presently. This paper introduces the image data discovery model IMDFSSM based on Hilbert Space Theory which uses pattern (defined as the vector in Hilbert Space) to represent the knowledge of image data quantitatively and participate in the knowledge discovery course. Then, it is validated by image mining system, and the results indicate that IMDFSSM can guide the knowledge discovery of image data.

Keywords Data mining, Image mining, Pattern

1 引言

最近几年,知识发现研究取得了很大的进展。目前,知识发现的应用对象从结构化数据发展到半结构化及非结构化的复杂类型数据,包括多媒体数据、图形图像数据及音频和视频数据等。知识发现开始克服传统分析方法的不足,适应于处理大规模信息量、复杂的以及具有非线性特征的数据类型,实现各类知识和信息的获取^[1]。

国内外知识发现的研究主要集中在知识发现的任务描述、知识评价与知识呈现,以基于各种理论的有效知识发现算法^[2,3]。目前,知识发现机理和一般性框架的理论研究日益受到重视,图像数据知识发现形成了新的研究热点。

本文着重探讨了基于 Hilbert 空间理论的图像挖掘系统的结构模型,同时用于遥感图像关联规则挖掘作为该总体结构模型的应用实例,以验证其有效性。

2 图像知识发现的理论基础

图像挖掘是指从图像数据集中抽取隐含的、以前未知的、具有潜在应用价值的、用户感兴趣的、与图像内容和空间关联和非关联的模式,它包括图像之间的关系、图像与字符数据之间的关系、图像中各实体之间的相互关系以及其他模式或关系等的过程。图像知识发现研究涉及到的理论基础有 Hilbert 空间和子空间理论等。

2.1 Hilbert 空间理论

我们采用 Hilbert 空间作为描述文本、图像集合等的状态空间,主要是因为:(1)Hilbert 空间适于描述具有复杂性的图像集空间;(2)Hilbert 具有正交基和正交分解等良好的性质。因此,Hilbert 空间更适于用来表征和分析处理复杂数据类

型。

定义 1 如果内积空间 X 作为导出范数下的一个赋范线性空间是完备的,则称 X 为 Hilbert 空间。

定义 2 设 $(H, \langle \cdot, \cdot \rangle)$ 是一个内积空间, $x, y \in H$, 若 $\langle x, y \rangle = 0$, 则称 x 与 y 正交, $x \perp y$ 。设 M 是 H 的一个非空子集, $x \in H$, 若 $\forall y \in M$ 都有 $x \perp y$, 称 x 与 M 正交。

定理 1 设 C 是 Hilbert 空间 H 的闭凸子集(凸是指 $\forall x, y \in C, \lambda \in (0, 1)$ 有 $\lambda x + (1 - \lambda)y \in C$) 则在 C 上存在唯一的最佳逼近元 x_0 , 使得 $\|x_0\| = \min_{x \in C} \|x\|$ 。

定理 2 设 M 是 Hilbert 空间 H 的一个完备线性子空间, 则对 $\forall x \in H$ 存在唯一的 $y \in M$ 使得 $y = P_M x$, 且 $\|x - y\| = \min_{z \in M} \|x - z\|$, 即 y 是 M 中距 x 最近的点(投影定理)。

定理 3 设 M 是 Hilbert 空间 H 的一个线性闭子空间, 则对 $\forall x \in H$ 存在下列唯一的分解 $x = x_0 + y$, 其中 $x_0 \in M, y \in M^\perp$ (正交分解定理)。

定义 3 设 M 是 Hilbert 空间 H 的一个非空子集, 如果 M 中任意两个矢量都是正交的, 则称 M 为 H 的一个正交集, 若 M 中的每个矢量的范数均为 1, 则 M 为规范正交集。

2.2 子空间概述

由 Hilbert 空间的定义可知 Hilbert 空间是一个完备的线性赋范空间, 所以它必然是一个线性空间。在线性空间中存在线性变换, 通过线性变换可以构建子空间, 并可以利用子空间来对原始空间进行描述和逼近。由于文本、图像等特征项空间是一个高维空间, 因此我们可以通过空间变换和映射将其分解到不同的子空间(概念空间)中对其进行分析和研究。

定义 4 设 M 是数域 K 上线性空间 H 的非空子集, 如果 M 对于 H 上的线性运算也构成数域 K 上的线性空间, 则

^{*}) 获国家自然科学基金重点项目(69835001)、国家教育部科技重点项目(教技司[2000]175)资助。杨炳儒 教授, 博士生导师, 主要研究方向: 推理机制与知识发现, 柔性建模与集成技术。游福成 博士, 副教授, 主要研究方向: 知识发现。

称 M 为 H 的一个线性子空间(简称子空间)。

定理 4 数域 K 上线性空间 H 的非空子集 M 是 H 的一个线性子空间,当且仅当 M 对于 H 的两种运算封闭,即:(1)若 $\alpha, \beta \in M$, 则 $\alpha + \beta \in M$; (2)若 $k \in K, \alpha \in M$, 则 $k\alpha \in M$ 。

定理 5 若 M 是有限维线性空间 H 的子空间,则 M 的一组基可以扩充成 H 的一组基。

设 $\alpha_1, \alpha_2, \dots, \alpha_s$ 是数域 K 上线性空间 H 的一组向量,这个向量组的所有线性组合构成的集合 W , 即:

$$W = \{k_1\alpha_1 + k_2\alpha_2 + \dots + k_s\alpha_s \mid k_i \in K, i=1, \dots, s\}$$

显然, W 是 H 的非空子集, 并且是 H 的子空间, 我们称 W 为由向量 $\alpha_1, \alpha_2, \dots, \alpha_s$ 张成的子空间, 记为: $\text{span}(\alpha_1, \alpha_2, \dots, \alpha_s)$ 或 $L = L(\alpha_1, \alpha_2, \dots, \alpha_s)$ 。所张成子空间的线性无关向量的数目 s 为子空间 L 的维数, $s = \dim(L)$ 。

定理 6 设 $V_1, V_2, \dots, V_n$ 是数域 K 上线性空间 H 的子空间, 如果 $V_1 + V_2 + \dots + V_n$ 中的每个向量 a 都可以唯一地表示成 $a = a_1 + a_2 + \dots + a_n, a_i \in V_i$, 则称 $V_1 + V_2 + \dots + V_n$ 为直和, 且有 $\dim(V_1 + V_2 + \dots + V_n) = \dim(V_1) + \dim(V_2) + \dots + \dim(V_n)$ 。

从空间变换的角度来看, 文本、图像挖掘的过程中存在许多变换, 例如降维、特征抽取、分类、聚类等都可以看作是空间的映射与变换。线性空间的映射是借助于矩阵运算来完成的。

定义 5 设 V_1 与 V_2 是数域 K 的两个线性空间, T 是 V_1 到 V_2 的一个映射, 若 V_1 中的任意两个向量 α, β 和 $\forall k \in K$, 都有

$$T(\alpha + \beta) = T(\alpha) + T(\beta), T(k\alpha) = kT(\alpha)$$

则称 T 是 V_1 到 V_2 的线性映射。线性空间 V_1 到 V_2 的映射, 也称为同构映射。

定义 6 设 V 是数域 K 上线性空间, V 到自身的线性映射称为 V 上的线性变换。线性变换是一类特殊的线性映射, 所以有关线性映射的性质对于线性变换都适应。

2.3 Hilbert 空间的逼近方式

经过实验分析表明, 原始文本、图像空间的表征维数很高(可达上万维), 我们可以采用特征抽取、映射以及空间近似(即函数空间中的逼近方法)等方法进行空间降维。其中空间近似是函数空间中的最佳逼近问题。在 Hilbert 空间中存在最佳逼近、具有最佳逼近的唯一性, 同时有着较好的算法来实现最佳逼近。

设 $(H, \|\cdot\|)$ 为线性赋范空间, $T = \{g_1, g_2, \dots, g_n\} \subset H$ 为空间 H 中的一组线性无关向量, 且 $V = \text{span}(T)$ 。空间近似问题可表述为: 对于空间向量 $X \in H$, 寻找系数 $c_1, c_2, \dots, c_n$ 使得 $\hat{X} = c_1g_1 + c_2g_2 + \dots + c_ng_n$ 与 X 尽可能地接近, 即

$$X = \hat{X} + e = c_1g_1 + c_2g_2 + \dots + c_ng_n + e$$

其中, e 是近似误差。如果 $X \in V$, 则可以找到这些系数。

从逼近论的观点来看, 对于极小化的(最小平方)逼近问题, 内积空间中的规范正交向量集有一个较好的性质, 即规范正交集能够使我们按照下述定理给出问题的显式解, 从而使问题易于求解。

定理 7 在 Hilbert 空间中, 逼近问题表述为: 若 $\{g_1, \dots, g_n\}$ 表示内积空间 H 中的规范正交集, 其范数定义为 $\|h\| = \sqrt{\langle h, h \rangle}$, 则表达式 $\|\sum c_i g_i - X\|$ 为最小当且仅当 $c_i = \langle X, g_i \rangle$ 。

可以证明定理中的系数 c_i 的唯一性^[4]。同时根据 Hilbert 空间投影定理可知, 若 V 是内积空间 H 中的一个有限维子空间, 则 H 中的每个向量都在 V 中有唯一的最佳逼近 $\hat{X}$, 即用唯一的广义多项式 $\sum c_i g_i$ 在最小平方意义下逼近一个给

定的向量:

$$X = \sum_{i=1}^n c_i g_i + e = \hat{X} + e$$

3 基于 Hilbert 空间理论的图像知识发现模型

基于以上提出的理论基础及图像数据的模式表示方法, 我们给出如下的图像知识发现模型 IMDFSSM, 见图 1 所示。

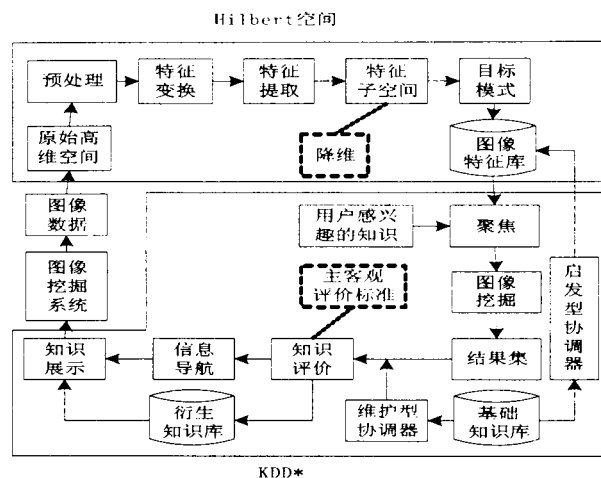


图 1 图像知识发现模型 IMDFSSM

3.1 图像数据的知识表示方法——模式表示

在人的认知过程中模式起着重要的作用。模式作为知识表示方法和心理表征方法被广泛地应用于知觉、注意、表象、记忆、思维解释、学习等心智过程。这些过程是模式的映射和变换过程。正是利用模式, 人类可以表示客观事物的复杂结构和性态, 利用模式使得认知过程变得有效而且易于理解。

为了便于图像数据的知识发现的研究, 寻找一种有效的知识表征方式, 本文从知识发现的角度给出模式形式化定义^[5]。

定义 7 在有限维 Hilbert 实空间 R^n 内, 可以定义一组线性无关的正交向量集所张成的子空间内的空间向量 $P_m = (u_1, u_2, \dots, u_j, \dots, u_m)$ 为模式。 u_j 为某个时空状态中模式的分量, 可分别表示模式的特征、性质、拓扑结构、组成关系等。线性无关的向量数目 m 为 P_m 的维数。

模式是图像挖掘过程中的一种知识表征方式, 是图像挖掘过程中的基本运算单元, 是具体或抽象的图像数据的量化描述。模式将参与图像知识的发现过程并表征所获得的知识。

定义 8 如果两个维数相同的模式的特征分量仅有微小差别, 则称两个模式具有相似性。在模式的分类、聚类过程中, 相似性是通过距离或其他测度来定义的。相似性常用向量之间的内积或用夹角余弦值来表示。

定义 9(模式映射) 在 Hilbert 空间中, $P = (u_1, u_2, \dots, u_m)$, $Q = (v_1, v_2, \dots, v_n)$ 为两个模式子空间, T 是一个法则, P 中的每个元素 u_i 通过 T 在 Q 中存在唯一的一个元素 v_j 与之对应, 称 T 为定义在 P 上, 取值在 Q 上的模式映射 $T: P \rightarrow Q$, 其中 T 为广义映射。根据实际需要 T 可为满足不同条件的映射法则, 如线性变换、拓扑变换、压缩映射乃至更为复杂和抽象的变换。

定义 10(模式分解) 对于某个具体的对象来说, 用已有模式作为一个整体来进行操作很困难, 即该模式为复杂模式, 则应将其划分为若干相对独立的由一定关系相联系的子模式。子模式分别表示所研究对象的不同的抽象层次和组成部

分。

通过在高维的原始空间进行特征抽取,形成原始图像数据集,然后在此基础上进行特征变换,构造维数适中的图像特征子空间,在该图像特征子空间可以利用矩阵的奇异值分解变化和近似计算方法来构造模式。最终用模式来表征图像数据,使得后续的各种处理过程可以参考结构化知识发现过程。

3.2 复杂类型数据的知识发现过程

基于模式的图像知识发现过程是一个发现新模式或对模式进行某种确证的过程。由于模式定义在 Hilbert 空间中,因而基于模式的图像知识发现同空间变换紧密地联系在一起。可同分类、聚类、相似模式等收敛性的知识发现算法及预测、时序等发散性的知识发现算法相结合来完成各种类型的图像知识发现。

图 1 中涉及了双库协同机制,这里我们给出双库协同机制非形式化的描述,详细内容参见文[6]。在给定真实数据库和基础知识库的前提下,在数据挖掘过程中,我们称具备以下特征的 KDD 中的运行机制为“双库协同机制”:

(1)数据库上,按数据子类结构形式所构成的挖掘数据库的可达范畴与基于属性间关系的挖掘知识库的推理范畴之间建立等价关系;两个范畴的等价关系为定向挖掘和定向搜索奠定了理论基础。

(2)在 KDD 聚焦过程中,除依据用户需求确定聚焦外,通过启发协调算法可以形成依挖掘知识库中知识短缺而生成的机器自身提供的聚焦方向,进而形成在数据库中的定向挖掘(算法和进程)。

(3)在获得假设规则到知识评价的过程中产生中断进程,即先不对假设规则进行评价,而是通过中断协调算法到挖掘知识库中进行定向搜索(算法和进程),以期发现产生的假设规则与知识库中原有的知识是否重复、冗余和矛盾,并作相应处理,即对知识库进行实时维护。

3.3 模式的评价

在这里,我们将构造模式评价函数,同时结合评价的主客观标准,采用定量的方式来评估结果模式集中有效的、新颖的、潜在可用的及最终可理解的模式,并把它存放到知识库中。

3.4 模式的解释与呈现

由于模式本身的可视性不强,在此我们将结合超图模型来表示相应的模式,用图形的方式直观地反映模式集。超图模型不仅可以形象化的方式来表示知识结构,简化复杂的知识结构,使得领域专家通过可视化途径进行模式的操作;同时模式的超图模型同面向对象技术有着很好的对应关系,易于采用面向对象技术编程实现模式的可视化。

4 实例验证——图像挖掘

这里主要介绍图像关联规则挖掘,其过程包括:特征变换、特征空间降维、特征提取、特征子空间形成、图像特征库 IFD 的构建、在特征库的基础上进行关联规则的挖掘等过程。

图像数据来源于文[7]提供的遥感图像数据。该文献通过概念提升,得到了 4 种属性特征数据:植被覆盖度(Coverage)、坡度(Slope)、耕地分布(Farm)和土壤侵蚀强度(Eros),由它们张成特征子空间,每一组值就对应着特征子空间中的一个点。

4 个属性分割之后,得到 10 个原始项,分别是 Coverage-low(植被覆盖率低 $<30\%$), Coverage-mid(植被覆盖率中等 $>30\%$), Coverage-high(植被覆盖率高 $>60\%$), Slope-low(坡度小 $<25^\circ$), Slope-mid(坡度中等 $<25^\circ$), Slope-high(坡度大

$>75^\circ$), Farm-true(耕地), Farm-false(非耕地), Eros-low(轻度侵蚀), Eros-high(强侵蚀)。

使用关联规则挖掘算法进行挖掘,部分挖掘结果如图 2、图 3 所示。通过关联规则挖掘,得到了一些重要的知识,并可以用于决策:

(1)坡耕地一般导致强度侵蚀(规则 5),非耕地一般为轻度侵蚀(规则 1);坡度在 25° 以下的地区大部分侵蚀强度较轻(规则 2);植被覆盖度在 30% 以上绝大部分($>94.3\%$)不会产生强度侵蚀(规则 3,4)。

(2)非耕地只有在低植被覆盖($<30\%$)和高坡度($>50^\circ$)的情况下才会与高侵蚀产生强关联(规则 18);耕地 $\rightarrow$ 低侵蚀的必要条件为坡度 $<25^\circ$ (规则 14,23);低植被覆盖区域一般产生强侵蚀,但如果坡度 $<25^\circ$,平均侵蚀强度将减弱(规则 8,17,23)。

序号	条件1	条件2	条件3	结果	支持度(%)	可信度(%)
5	farm-true			eros-high	5.7	71.5
1	farm-false			eros-low	60.3	64.1
2	slope-low			eros-low	19.5	81.7
3	coverage-mid			eros-low	13.1	94.3
4	coverage-high			eros-low	3.2	100
6	farm-false	slope-low		eros-low	20.1	62.3
7	farm-false	coverage-mid		eros-low	11.5	96.1
8	coverage-low	slope-low		eros-low	15.4	77.9
9	coverage-low	slope-high		eros-high	5.1	66.5
10	coverage-low	farm-true		eros-high	4.22	71.2
11	slope-mid	coverage-mid		eros-low	8.5	95.4

图 2 图像关联规则挖掘(1)

序号	条件1	条件2	条件3	结果	支持度(%)	可信度(%)
14	slope-low	farm-true		eros-low	2.1	72.7
15	coverage-mid	slope-high		eros-low	0.91	76.12
16	slope-high	farm-false		eros-high	0.86	99.2
17	farm-false	coverage-low	slope-low	eros-low	16.21	76.82
18	farm-false	coverage-high	slope-high	eros-high	4.15	51.56
19	farm-false	coverage-mid	slope-mid	eros-low	8.21	96.21
20	farm-false	coverage-mid	slope-high	eros-low	3.25	95.75
21	farm-false	coverage-low	slope-high	eros-low	0.782	75.3
22	farm-true	coverage-low	slope-mid	eros-high	3.21	65.2
23	farm-true	coverage-low	slope-low	eros-low	1.12	74.23
24	farm-true	coverage-high	slope-high	eros-high	0.46	95.4

图 3 图像关联规则挖掘(2)

因此,为了有效防治研究区内水土流失和搞好生态环境建设,应采取两个重要的措施:(1)增加地表植被的覆盖度(争取达到 30%);(2)坡度 $>25^\circ$ 区域严禁开垦,已有的坡耕地坚决退耕还林、还草。

结论 在探讨了基于 Hilbert 空间理论的图像知识发现模型之后,我们用图像挖掘系统作为实例进行了验证,结果表明模型具有较好的实用性与有效性。

参考文献

1. Piatetsky-Shapiro G, Frawley W J. Knowledge Discovery in Databases[C]. AAAI/MIT Press, 1991. 166~175
2. Fayyad U M, Piatetsky-Shapiro G, Smyth P, Uthurusamy R. Advances in Knowledge Discovery and Data Mining[C]. AAAI/MIT Press, 1996. 20~32
3. Fayyad, et al. eds. Advances in Knowledge Discovery and Data Mining[J]. AAAI/MIT Press, 1996. 1~35
4. [美]E. W. 切尼. 逼近论导引[M]. 上海科学技术出版社, 1981
5. 张德政. 基于相似模式知识发现方法的研究与应用[D]. 北京科技大学, 2002
6. 杨炳儒. KDD 中双库协同机制的研究(I)[J]. 中国工程科学, 2002, 4(4): 26~32
7. 马超飞, 刘建强. 遥感图像多维量化关联规则挖掘[J]. 遥感技术与应用, 2003, 18(4): 243~247