数据挖掘方法本体研究*)

邹力鹏[1] 王丽珍[1] 姚绍文[2]
(云南大学信息学院计算机科学与工程系 昆明650091)[1]
(云南大学软件学院 昆明650091)[2]

摘 要 数据挖掘是包含多个阶段的知识发现过程。一个简单、但典型的数据挖掘过程可能包括数据预处理阶段,数据挖掘算法的应用阶段,以及对挖掘结果的可视化处理阶段。在每个阶段,都会有多个算法或方法供数据挖掘工作者选择,但仅有一些算法和方法组合是有效的。即使是数据挖掘领域的专家,也可能会忽略一些重要的、有助于知识发现的数据挖掘算法或方法。本文中,我们将讨论使用本体的方法来协助数据挖掘工作者在实施数据挖掘过程中对众多可供选择的算法和方法进行选择。

关键词 数据挖掘,知识发现,本体

Research on Ontology of Data Mining Methods

ZOU Li-Kun[1] WANG Li-Zhen[1] YAO Shao-Wen[2]
(Department of Computer Science and Engineering, College of Information, Yunnan University, Kunming 650091)[1]
(College of Software, Yunnan University, Kunming 650091)[2]

Abstract A data mining (DM) process involves multiple stages. A simple, but typical, process might include preprocessing data, applying a data-mining algorithm, and visualizing the mining results. There are many possible choices for each stage, and only some combinations are valid. Therefore, both novices and data-mining specicalists could overlook some important, potentially fruitful optionos. In this paper, we introduce how to create the ontology of data mining methods in order to facilitate the choice of DM processes to execute.

Keywords Data mining, KDD, Ontology

1 引言

"本体论"最早是哲学中的基本概念,它是研究"是"之所以为"是"的理论,可以说是哲学中的哲学,甚至可以认为西方哲学自身的发展就是一个"本体论"的产生、发展、怀疑和批判的过程。近年来,本体论的方法在知识工程领域得到了越来越广泛的应用,在很多有名的知识系统中,如美国 D. Lenat 教授领导研制的大型常识知识库系统 Cyc,Princeton 大学 Berkeley 分校研制的语言知识库 WordNet 等,本体论都有一定的应用[1,2]。一方面,本体论研究深层次上的指示,把知识工程研究中的知识向更深更本质的方向上推进,另一方面,本体论的研究独立于任何语言,因此本体论将会为不同系统之间知识的共享和互操作提供手段。

早在1998年,Gruber 就已经给出了本体的一个流行定义,即"本体是领域概念化对象的明确表示和描述"。Guarino 把概念化对象 C 定义为:$C=\langle D,W,R\rangle$,其中 D 是一个领域,W 是该领域中相关的事务状态集合,R 是领域空间$\langle D,W\rangle$概念关系的集合[3]。因此,从概念化对象的定义来看,本体把现实世界中的某个领域抽象成一组概念(如实体、属性、进程等)及概念间的关系。某个领域的本体不仅提供了关于该领域的一个公认的概念集,同时也表达了各概念间所具有的各种语义联系[4]。

随着数据挖掘技术在商业领域中得到越来越广泛的应用,对数据挖掘算法以及方法的研究也日新月异,有关数据挖掘过程各阶段的新思想、新算法、新技术层出不穷。一个简单,但典型的数据挖掘过程可能包括数据预处理阶段,数据挖掘算法的应用阶段,以及对挖掘结果可视化处理阶段[5]。由于数据挖掘是包含多个阶段的知识发现过程,而在每个阶段,都会有多个算法或方法供数据挖掘工作者选择,但仅有一些算法和方法组合是有效的。因此,即使是数据挖掘领域的专家,在一个具体的挖掘任务进行到某一个阶段时,也难免会产生困惑:该阶段可用的技术有哪些?这些现成的技术是否合适?若不合适,采用的新技术以后能否被其他研究者使用?产生的结果是不是用户最需要的?

基于上述原因,受文[6]的启发,在本文中我们将本体的概念引入到数据挖掘方法中,不同于其他基于本体论的数据挖掘方法使用本体来表示领域知识,我们是为已经存在的、被证明可以有效使用的数据挖掘技术建立本体。通过数据挖掘方法本体,协助不论是数据挖掘领域的新手还是专家在实施数据挖掘过程中对众多可供选择的算法和方法进行选择。

2 基于本体的数据挖掘过程

基于本体的数据挖掘过程如图1所示。

首先,数据挖掘工作者通过和用户交流获得关于要挖掘的数据、终极目标以及什么是用户最急需知道的相关信息。第二步,根据用户输入所隐含的约束,数据的特点和已经存在的

*)基金项目:国家自然科学基金资助项目(60363006),云南省自然科学基金资助项目(2002F0013M)。邹力鹏 硕士,助教,主要从事数据挖掘研究。王丽珍 教授,主要从事数据挖掘及数据仓库方向研究。

本体获得所有有效的 DM 过程集合(每一个有效 DM 过程就是一个可实施的执行方案)。在这一步中包括如何选择适当的数据预处理、如何选择合适的数据挖掘算法和对挖掘结果进行优化、可视化的模型的操作。接着,根据用户所急需的需求顺序对第二步所得到的执行方案集合进行排列,形成一个可执行方案的清单,这样,用户可以在该方案清单中选择合适的执行方案。最后一步,选择清单中的计划并执行。

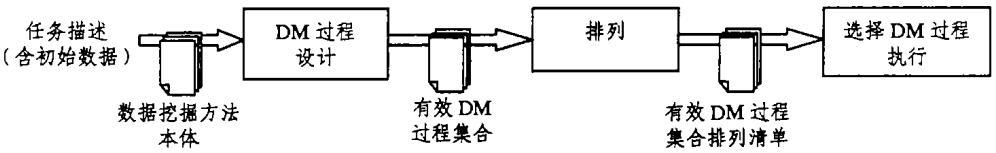

图1 基于本体的数据挖掘过程

在整个过程中,本体的作用就是使数据挖掘工作者不会忽略或遗忘那些他们所不熟悉、但又有可能导致发现知识的数据挖掘技术。

是不是在第二步有了可选方案,整个数据挖掘的工作就结束了?答案是否定的,有了方案并不意味着工作的结束,我们还应该考虑"什么对用户来说是重要的?"。

例如,图2中显示了执行分类任务的3个不完全统计方案,得到模型之后,仍然需要优化和可视化处理才能得到所需要的分类知识。方案1中仅仅包含了判定树的引导器,方案2中:首先是对数值型数据进行离散化处理,然后再建立朴素贝叶斯分类器。方案3中:首先对数据进行随机抽取子样品,第二步进行离散化处理,最后建立朴素贝叶斯分类器。

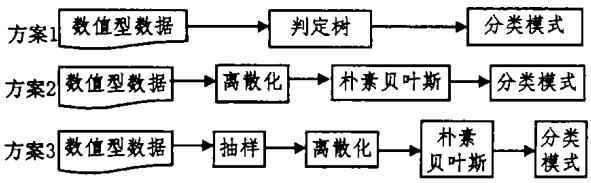

图2 执行分类任务的3个方案

对于不同的用户而言,三个方案的意义可能是不一样的。如果用户希望尽可能地减少操作步骤,则在图2中,用户会优先考虑方案1,接下来是方案2,最后才是方案3。但如果用户会希望运行的时间尽可能的短,以期能尽快地得到结果,则用户会优先考虑方案3,接下来是方案2,最后是方案1。

换言之,对不同的用户,可执行的方案计划按照用户的需求可能会有不同的排列次序。为了让最终的结果符合用户的需求,有了可选方案是不够的,还需要对第二步得到的执行方案集合按照用户的需求进行排列。

3 数据挖掘方法本体的建立

在挖掘过程中,本体是用来协助用户构成有效的 DM 过程(可执行方案)集合。数据挖掘方法本体需要定义已经存在的数据挖掘技术以及其特征属性。首先,我们对每个操作(算法或过程)建立本体,本体中包含如下信息:

- 每个操作的可读信息;
- 对于每个操作,说明其执行环境,包括前提条件以及该操作和前驱操作的兼容性;
- 操作执行的结果的详细说明;
- 说明阈值情况;
- 对影响操作属性如速度、精度、模型复杂性的估计。

APRIOR 算法、朴素贝叶斯的本体描述如图3所示。

将所有方法本体综合在一起就可以构成数据挖掘方法的本体。本体对所有操作按照逻辑的形式进行分类,形成不同的逻辑组。在数据挖掘工作进行到某个阶段时,这些逻辑组可以用来减少构成有效 DM 过程的操作的数量。

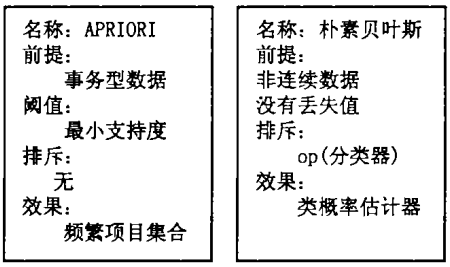

图3 APRIOR 算法、朴素贝叶斯的本体描述

一个有效的 DM 过程并不会违背本体中任何叶节点的基本约束。例如,如果输入的数据集合包含数值属性的数据,则将数据简单地应用于朴素贝叶斯算法是不合理的,因为朴素贝叶斯仅仅能处理分类属性。此时可以使用离散化规则预处理数据使数据能够完成从数值型到分类属性的转换,再使用朴素贝叶斯。

数据挖掘方法本体建立后,还可以为彼此工作相互独立的挖掘工作组或工作者提供共享新成果的平台。例如,当工作组 A 和工作组 B 彼此独立工作时,工作组 A 若发现现有的技术都不能较好地实现用户需求时,可以设计出新的算法或过程,只需将它加入到我们的本体系统中,则不论是工作组 B 还是其他的工作者以后都可以使用该新技术。这也是我们提出为数据挖掘方法建立本体的初衷之一,即如何实现数据挖掘工作者之间的信息共享,使彼此的工作不再处于独立状态。

4 算法设计

在本节中,我们将给出使用数据挖掘方法本体生成有效 DM 过程的算法伪码。算法设计思想如下:我们通过生成一棵树来生成所有可执行方案。初始时,生成树只有一个根节点,在该节点中保存有初始数据特点。同时根据目标任务描述将叶节点的最小子类按照从左到右的顺序排列形成一个子类的有序集合(目标任务描述的作用是在挖掘算法本体中将彼此独立、不可能连续执行的技术排除,例如:当用户要完成关联分析任务时,则分类、偏差检测等本体将不会被包含在该集合中),每一个子类中有若干本体。我们对生成树的所有叶节点按照广度优先的次序完成如下操作:根据当前叶节点的数据特点,在第一个子类中所包含的方法本体中查找前提与之相一致,不存在操作排斥性的技术,有则生成一个新节点,在该节点中记录三方面信息:技术的名称,排斥信息以及执行完该技术后数据的特征,即方法本体中所描述的效果。新节点作为当前叶节点的子节点插入,对于同一个当前叶节点,可能会生成有多个子节点,同时,还需要插入一个节点,该节点在技术

名称位置为空，表示没有采用该本体中的技术，其数据特点是当前节点的数据特征。当处理完生成树同一层次中最后一个叶节点时，表明在该阶段的所有可能的技术组合都已经考虑，可以进行下一阶段操作。既考虑子类集合中的下一个子类，直至将集合中的所有子类都遍历之后，我们也将生成相应的生成树了。

```
输入：只有根节点的树 T，子类的有序集合{C1,C2,…,Cn}
输出：有效 DM 过程的生成树
过程：
for (i=1,i<=n,i++)
{ leaves=getleaf(T);//得到 T 的叶子节点集合；
  for each r in leaves
  { for each ontology Oj in Ci
    {if ( r.数据特征= Oj.前提 and r.排斥() Oj.名称)
       new(t,Oj.名称,Oj.效果,Oj.排斥);
     //生成新节点 t,名称为 Oj 所代表的技术,数据特征为执行该技
           术后的结果 add(t,r);//将 t 作为 r 的子节点加入 T 中
    }
    new(t,null,r.数据特征,r.排斥);
    //生成一个没有名称的新节点，表示没有采用 Ci 中的任何技术。
    add(t,r);
  }
}
```

我们对 T 进行遍历生成所有最长路径，一条最长路径上的所有节点，即为一个可执行计划方案中所有细节。

总结 数据挖掘是一个由多个阶段组成的知识发现过程，在每个阶段都有很多的相关技术。随着数据挖掘技术在商业领域中的日益普及，越来越多的新技术被提了出来，此时，不论是数据挖掘领域的专家还是新手，都可能会忽略有用的技术。为此，我们提出为数据挖掘方法建立本体，来解决上述问题，并初步建立数据挖掘方法的本体以及相关算法。更重要的是，数据挖掘方法本体的建立还可以为数据挖掘工作者之间共享信息提供平台，使他们的工作彼此不再独立。

本文中，我们在概念上探讨将本体引入数据挖掘方法中，并对数据挖掘方法本体和其相关算法进行了初步设计，目的在于帮助数据挖掘工作者在工作过程中，面对如何选择数据挖掘技术时不再困惑。目前，我们已经在动手建立部分本体，以实现本文中所提出的算法，同时在着手设计对得到的方案计划按用户需求进行排列的算法。

下一步的目标是考虑如何更好地共享知识发现成果，实现所谓的网络外延性，我们将基于课题项目设计一个原型系统。

参 考 文 献

1 陆汝泠. 世纪之交的知识工程与知识科学. 清华大学出版社，2001
2 周肖彬，曹存根. 基于本体的医学知识获取. 计算机科学，2003，30(10)：35～39
3 Uschold M，Gruninger M. ONTOLOGIES：Principles，methods and applications. Knowledge Engineering Review，1996，11(2)：93～155
4 Guarino N. Formal ontology and information system. In：Guarino N ed. Formal Ontology in Information System. Trento：IOS Press，1998. 6～8
5 Han Jiawei，Kambr M. Data Mining Concepts and Techniques. 高等教育出版社，2001
6 Bernstein A，Provost F，Hill S. An Intelligent Assistant for the Knowledge Discovery Process：An Ontology-based Approach. 2002. Working Paper of the Center for Digital Economy Research，New York University - Leonard Stern School of Business，CeDER Working Paper # IS-01-01
7 Keim D A. Information Visualization and Visual Data Mining. IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS，JANUARY-MARCH，2002，7(1)：100～107

（上接第189页）

12 Kononenko I，Simec E，Robnik-Sikonja M. Overcoming the myopic of inductive learning algorithms with RELIEFF. Applied Intelligence，1997，7(1)：39～55
13 Koller D，Sahami M. Toward optimal feature selection. In：Proc. 13th Int. Conf. Machine Learning，Morgan Kaufmann，1996. 284～292
14 Chow C，Liu C. Approximating discrete probability distributions with dependence trees. IEEE Transactions on Information Theory，1968，14：462～467
15 Berger A L，Della Pietra S A，Della Pietra V J. A maximum entropy approach to natural language processing. Computational Linguistics，1996，22(1)：39～72
16 Bender O，Josef Och F，Ney H. Maximum Entropy Models for Named Entity Recognition. In：Proc. of the Seventh CoNLL conf. Edmonton，May-June 2003
17 Chieu H L，Ng H T. Named Entity Recognition with a Maximum Entropy Approach. In：Proc. of the Seventh CoNLL conf. Edmonton，May-June 2003. 160～163
18 Freitag D，McCallum A. Information Extraction with HMM Structures Learned by Stochastic Optimization. In：Proc. of AAAI-2000
19 Freitag D，MaCallum A K. Information Extraction with HMMs and Shrinkage. AAAI99
20 Rabiner L R. A tutorial on hidden Markov models and selected applications in speech recognition. Proc. of the IEEE，1989，77(2)
21 McCallum A，Freitag D，Pereira F. Maximum entropy Markov models for information extraction and segmentation. In：Proc. ICML，Stanford，California，2000. 591～598
22 Lafferty J，McCallum A，Pereira F. Conditional random fields：Probabilistic models for segmenting and labeling sequence data. In：Proc. ICML
23 McCallum A，Li Wei. Early Results for Named Entity Recognition with Conditional Random Fields，Feature Induction and Web-Enhanced Lexicons. In：Proc. of the Seventh CoNLL conf. Edmonton，May-June 2003
24 Sha F，Pereira F. Shallow parsing with conditional random fields. In：Proc. of Human Language Technology，NAACL
25 Furey T，Cristianini N，Duffy N，Bednarski D，Schummer M，Haussler D. Support vector machine classification and validation of cancer tissue samples using microarray expression. Bioinformatics，2000
26 Lodhi H，Saunders C，Shawe-Taylor J，Cristianini N，Watkins C. Text classification using string kernels. Journal of Machine Learning Research，2002
27 Collins M，Duffy N. Convolution kernels for natural language. In：Proc. of NIPS-2001，2001
28 Zelenko D，Aone C，Richardella A. Kernel methods for relation extraction. Journal of Machine Learning Research，2003
29 Dietterich T G. An experimental comparison of three methods for constructing ensembles of decision trees：Bagging，boosting，and randomization. Machine Learning，2000，40(2)：139～157
30 Ghani R. Using error-correcting codes for text classification. In：P. Langley，ed. Proc. of ICML-00，17th Intl. Conf. on Machine Learning，Stanford，US：Morgan Kaufmann Publishers，San Francisco，US，2000. 303～310
31 Florian R，Ittycheriah A. Named Entity Recognition through Classifier Combination. In：Proc. of the Seventh CoNLL Conf. Edmonton，May-June 2003. 168～171
32 Kleinberg E M. A Mathematically Rigorous Foundation for Supervised Learning. In：J. Kittler，F. Roli，eds. Multiple Classifier Systems. First International Workshop，MCS 2000，Cagliari，Italy，volume 1857 of Lecture Notes in Computer Science，Springer-Verlag，2000. 67～76
33 Allwein E L，Schapire R E，Singer Y. Reducing multiclass to binary：a unifying approach for margin classifiers. Journal of Machine Learning Research，2000，1：113～141
34 Avrim B，Mitchell T. Combining labeled and unlabeled data with co-training. In：Proc. of the Workshop on Computational Learning Theory. Morgan Kaufmann，1998
35 Collins M，Yoram S. Unsupervised models for named entity classification. In：Proc. of the 1999 Conf. on Empirical Methods in Natural Language Processing，1999
36 Nigam K，Ghani R. Analyzing the effectiveness and applicability of co-training[A]. In：Proc. of Ninth Intl Conf. on Information and Knowledge Management (CIKM)[C]，2000
37 Thompson C A，Califf M E，Mooney R J. Active Learning for Natural Language Parsingand Information Extraction. In：Proc. of the Sixteenth Intl. Machine Learning Conf. Bled，Slovenia，June 1999. 406～414
38 Muslea I，Minton S，Knoblock C A. elective sampling with redundant views. AAAI/IAAI
39 Jones R，Ghani R，Mitchell T，Riloff E. Active Learning with Multiple View Feature Sets. ECML 2003 Workshop on Adaptive Text Extraction and Mining