

摘要

随着数据库、信息系统、互联网的迅速发展，人们可获取的数据繁多。在这些数据背后隐藏着许多重要的信息，因此如何更好地分析并利用数据越来越受到人们的关注。偏差分析作为重要的数据分析工作的一部分，也成为研究的主要课题之一。

面向解释的偏差分析是以本体和问题解决方法为基础，针对互联网以及众多正在运行的信息系统，把数据现状和历史纪录或标准进行比较、找出偏差、分析偏差，得到人们所能理解的原因。它可以纳入知识工程中启发式分类的范畴。

本文作为面向解释的偏差分析的前半部分工作，主要是通过对我们所做的一些偏差分析系统进行比较、分析，将面向解释的偏差分析问题模型化，并提出相应的问题解决方法。在模型化问题时，针对一些应用领域没有提供具有适宜分类层次的领域的情况，采用属性和值分开的方法加以解决：通过应用解释标准在解释空间进行匹配；使用匹配标准在由解释属性构造的分类层次结构上搜索类别并在属性值表上区分影响程度，最终得到带有抽象类别的人易理解的偏差解释。这是提出问题解决方法的前提，也是后半部分工作的基础、构造本体的依据和前期准备工作。

面向解释的偏差分析给出了一个基于本体的偏差分析问题解决方法，它补充了启发式分类的类型，具有较好的实用性、时效性和可重用性，有助于资助和改善人们的解决问题能力和决策能力。因为我们的最终目标是实现用于知识处理的语义 Web 服务，所以希望今后将面向解释的偏差分析系统在网上作为语义 Web 服务中的一种服务进行发布，并在这个语义 Web 服务上添加其它的服务，从而更准确、更高效地解决实际应用中存在的大量问题，为各行各业的信息化建设做出贡献。

关键词 面向解释的偏差分析；启发式分类；解释标准；匹配标准；属性值表

Abstract

With the rapid development of database, information system and internet, people could obtain more and more data. Because the data contain a lot of important information, people get focus on how to analyze them and make use of them better. As a part of important task about data analysis, deviation analysis becomes one of the main subjects studied.

Aiming at internet and information systems, explanation-oriented deviation analysis based on ontology and problem solving methods compares the data with standards, finds out and analyzes the deviation to help people understand the reason. It could be put into the category of heuristic classification in knowledge engineering.

As the first half part work of this analysis, this paper models the problem and proposes the problem solving method by analyzing some accomplished systems about deviation analysis. When modeling the problem, aiming at some application fields without providing domain knowledge having appropriate classification hierarchy, this paper adopts the method of separating attributes and their values to solve it: applying the explanation standard to match in explanation space, using the matching standard to search the class in the hierarchy structure for classification constructed by explanation attributes and distinguish the influence degree in attribute value table, so we could finally obtain the intelligible deviation explanation having abstract class. This work is the precondition of proposing the problem solving method, the base of the second half part work, and the preparation of constructing the ontology.

Explanation-oriented deviation analysis presents a problem solving method for deviation analysis based on ontology. It extends the type of heuristic classification, has good practicality, timeliness and reusability, and helps people solve problems and make decision. Because our ultimate aim is to implement Semantic Web Services for knowledge processing, we hope to put our analysis system on the web as a service of it and add other services to it in future. So it will solve a lot of problems in practical application more accurately and efficiently, and make contribution to the information construction of all trades and professions.

Keywords explanation-oriented deviation analysis ; heuristic classification ; explanation standard; matching standard; attribute value table

独创性声明

本人声明所呈交的论文是我个人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得北京工业大学或其它教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

签名： 马玉莲 日期： 2007年6月6日

关于论文使用授权的说明

本人完全了解北京工业大学有关保留、使用学位论文的规定，即：学校有权保留送交论文的复印件，允许论文被查阅和借阅；学校可以公布论文的全部或部分内容，可以采用影印、缩印或其他复制手段保存论文。

（保密的论文在解密后应遵守此规定）

签名： 马玉莲 导师签名： 马玉莲 日期： 2007年6月6日

第1章 绪论

1.1 引言

随着微电子技术、计算机技术的发展,以计算机为工具的信息处理技术在经历了数值计算阶段、数据处理阶段后,已经进入知识处理的新阶段。因此以知识为处理对象的知识工程(Knowledge Engineering)应运而生。它作为一门应用学科,从诞生之日起就是人们研究的一个热门课题。早期知识工程的研究侧重于知识的表示及推理机制。但是目前知识工程进入了一个新时代,为了更好地解决问题,其重点在于实现知识的共享和重用。由于本体和问题解决方法的实质性作用都是支持重用,因此将二者结合起来进行研究成为知识工程发展的主流方向、有助于提高人们处理问题和解决问题的能力。

目前,Web的迅速普及为知识的共享和重用创造了条件。其主要的两个新兴发展趋势是:一个是Web服务,即一些发布在网上的可被调用的模块化程序,实质上就是问题解决方法;另一个是语义Web,即有着良好语义定义的Web,其核心思想是构造大量计算机可理解的领域本体。而语义Web服务的思想就是将Web服务和语义Web结合起来,从而提供一种更为智能的推理行为。为了人们能够更好地共享与重用网上的知识,语义Web服务作为一种本体与问题解决方法在Web上的结合,成为一个非常值得研究的问题。

此外,随着数据库、信息系统、互联网的迅速发展,人们可获取的数据越来越多、越来越复杂。在这些数据背后通常隐藏着许多重要的信息和潜在而有用的知识,因此如何对其进行高水平的分析和处理、从而更好地利用数据来指导人们的工作和学习越来越受到人们的关注。偏差分析作为非常重要的数据分析工作的一部分,也越来越受到重视,成为研究的主要课题之一。

基于以上原因,构建一个偏差分析的语义Web服务是非常重要而有意义的。目前,我们要做的就是构建偏差分析语义Web服务的基础工作,即:将本体和问题解决方法结合起来进行面向解释的偏差分析的研究。

1.2 研究背景与研究内容

对于偏差进行分析在数据分析工作中是很重要而有意义的,原因在于:这些偏差对分析的结果有很大的影响;有一些偏差在应用领域体现了一些有趣的现象或重要的信息;对于大多数应用来说,识别出来的偏差通常能导致意外的知识的发现^[1]。因此,分析数据中存在的偏差、并得到对于其产生原因的解释是十分必

要的。但是，目前一般的偏差分析工作主要是检测或寻找海量数据中出现的含有重要信息的异常点^[2, 3]，而寻找产生偏差的解释的工作在数据分析领域进行得很少，因此必须寻找一种及时有效的针对解释的偏差分析方法。面向解释的偏差分析便是基于这些方面的考虑由实际需要产生的，它符合由William J. Clancey提出的启发式分类模型^[4]。因此，面向解释的偏差分析工作应适合在知识工程领域中加以探讨，且有广阔的应用前景。

在知识工程领域，近些年关于本体论的研究成果逐渐增多，已成为关注的热点。本体论是一种先进的知识表示技术^[5]，研究实体的存在性和实体存在的本质。和面向对象方法适合表示纵向继承关系相比，本体论更加适合知识元之间横向关系的描述。本体的构造可以让计算机对信息处理更加方便，实现知识的共享和重用，有助于发现更有意义的信息，从而更准确、更有效地解决问题。但是，随着本体论研究的深入发展，一些问题也呈现出来，表现在虽然在一些领域开发了本体，但应用很有限，另一方面，本体开发基本是手工，缺少有效的开发工具^[6, 7]。

问题解决方法（problem solving methods）^[8]是构建知识库系统(KBS)的主要组件，它描述了一个通用的推理行为，这一推理行为可以重用在不同的应用领域上。因此，问题解决方法在知识获取和知识工程中扮演着重要角色。问题解决方法的思想最早源于启发式分类。启发式分类揭示了一种在推理模式中存在的普遍行为，是从知识密集型领域抽象出来的，对不同层次的知识进行分类、整理，并有条理地组织起来，使其便于利用、维护和重用^[9]。启发式分类包括抽象、匹配、细化(refinement)三种操作，其中抽象与细化过程都需要较多的领域知识，因此有较大的难度并需要较高的花费。

由于本体和问题解决方法的实质性作用都是支持重用，二者的结合可以推动这两个领域的发展，有助于实现知识库系统中知识的共享和重用，推动知识库系统的开发和实际应用^[10-12]。因此基于本体的问题解决方法及其相关的方法论研究一直引人关注^[13, 14]，而我们的研究方向和重点也是把二者放在一起研究、并关注于二者的集成。

我们的研究内容是，为了有效地解决知识工程中知识的共享和重用问题、帮助和改善人的解决问题能力和决策能力，我们从互联网和信息管理系统出发，充分利用其中已有的领域知识，进行本体和问题解决方法的研究，并基于二者构建用于知识处理的语义 Web 服务，可以在知识的基础上对其中的海量数据进行高水平的数据分析、数据校验和数据咨询等工作，从而实现一个利用已有知识进行智能化信息处理的过程。这主要包括三方面的工作：（1）将分布在网络上关系数据库中的数据自动地转化为知识库即本体的形式；（2）对这些以本体形式存在的知识进行处理；（3）将这些知识和模块化的程序用 Web 的语义标记方式描述并发布到 Web 上成为语义 Web 服务。

在上述工作中，第一方面的基础研究工作已经完成^[6]，即将关系数据库中的数据半自动化地转化为本体的形式。第二方面的工作是对第一方面工作所得到的知识进行不同的处理，如数据校验^[7]、数据分析等。而偏差分析作为数据分析的重要工作之一，属于第二方面的工作，是我们目前研究的一个重要方面，因为偏差中包括很多潜在而有用的知识。

我们目前所做的主要工作就是：根据实际偏差分析的需求，以本体和问题解决方法为基础，提出了面向解释的偏差分析。面向解释的偏差分析是针对互联网以及众多正在运行的信息管理系统，对于需要把数据现状和历史纪录或标准进行比较、找出其偏差、分析偏差产生原因的大量问题进行分析，得到人们所能理解的原因，资助和改善人们的解决问题能力和决策能力。为了达到这一目标，需完成两个部分的工作：（1）通过对我们所做的一些偏差分析系统进行比较、分析，将面向解释的偏差分析问题模型化，提出面向解释的偏差分析问题解决方法。（2）以本体的形式构造可重用的面向解释的偏差分析系统，并通过UPML框架和JESS开发工具来实现这个通用系统。这些工作是由我和另一位同学共同完成的。其中我主要是负责第一部分的工作，同时参与了第二部分的工作。

本篇论文的内容就是对第一部分工作的描述，通过一些基本概念来模型化面向解释的偏差分析问题，并通过对试卷成绩、超市经营数据和球赛统计数据三个具体领域进行分析来详细说明了模型化的过程，最终总结出面向解释的偏差分析问题解决方法。为了模型化面向解释的偏差分析问题、并给出问题解决方法，在三年中我所做的工作具体包括：

（1）分别对多个不同领域的实例数据进行面向解释的偏差分析的研究：首先是去除不合理的偏差情况，从而确定可能出现的各种偏差情况和大小；其次是寻找对于可能出现的偏差的原因解释，对解释进行规范化。

（2）设计并实现了其中一个面向解释的偏差分析系统（试卷成绩分析系统），参与了本科生所做的其余四个面向解释的偏差分析系统的设计与实现工作（超市经营数据分析系统、球赛统计数据系统、水污染数据分析系统和中医诊断数据分析系统）。

（3）通过对我们所做的其中三个有着相似特点的偏差分析系统（试卷成绩分析系统、超市经营数据分析系统和球赛统计数据系统）进行比较、分析，总结出面向解释的偏差分析的整个过程，归纳出进行分析时所依据的不同参照或基准，最终抽象出一个比较通用的对偏差分析问题的形式化描述（模型化）和问题解决方法。

在模型化面向解释的偏差分析问题时，本文采用不同于经典启发式分类的一些抽象方法给出了较为规范的抽象过程；采用属性和值分开的方法，通过应用解释标准在解释空间进行匹配；使用匹配标准在由解释属性构造的分类层次结构上

搜索类别并在属性值表上区分影响程度，完成细化过程，最终得到带有抽象类别的人易理解的偏差解释。这是提出问题解决方法的前提，也是第二部分工作的基础、构造本体的依据和前期准备工作。

我们的分析研究主要针对观测结果与模型预测值的偏差，兼顾对异常数据的偏差的处理，这和现有的偏差分析的研究重点有所不同。不仅如此，我们所做的面向解释的偏差分析研究是基于本体的，这减少了反复构造知识库的代价，使知识的共享与重用成为可能，有助于发现更加有意义的信息，从而更准确、更有效地解决问题。另外，与经典启发式分类方法需要较多领域知识不同，本文的研究领域是一些虽然问题并不是很复杂、但一般没有提供具有适宜分类层次的领域知识的领域，如：试卷成绩分析、超市经营数据分析、球赛统计数据分析和领域；但由于二者都有一个启发式的匹配过程，所以将本文的研究纳入知识工程中启发式分类的范畴；但是在具体的分析方式上，本文又有自己的特点。

我的毕业论文工作的成果是模型化面向解释的偏差分析问题、并总结出面向解释的偏差分析问题解决方法，这是另一位同学的工作依据。我和另一个同学的工作成果合起来可提供一个用于解决偏差分析问题的通用系统，完成对偏差进行分析的功能，最终可以得到带有抽象类别的人易理解的偏差解释。

1.3 相关领域的研究进展及成果

1.3.1 本体

1、本体的定义

本体 (Ontology, o 大写) 一词起源于哲学领域，是对本质和存在的一个系统的描述^[16]。在计算机领域中，本体 (ontology, o 小写) 可用来描述和说明某一领域的知识、知识共享和重用。直观地讲，本体是一个实体，是对某领域应用本体的方法分析、建模的结果，本体中包括某一领域中基本概念、概念之间的关系以及概念和关系的属性的定义^[16-18]。按照 Gruber 的定义：“本体是概念化的一个显式的规格说明”。Borst 对这个定义稍微做了一点修改，提出：“本体可定义为被共享的概念化的一个形式的规格说明”^[19,20]。

作为一种先进的知识表示技术，本体成为包括知识工程、自然语言处理和知识表示在内的人工智能研究团体的热门研究课题。本体提出的最初目标是实现知识的共享和重用，这也是本体的主要作用和研究本体的意义所在。

本体的设计与实现通常分为5个阶段：知识获取，概念化，形式化，实现和维护。在设计的过程中同时兼顾知识的共享和交换。亦即必须先学习某个领域的

知识，在掌握该知识的基础上，分析其组成、结构、层次及相互的关系，然后将其抽象至概念层次，使用形式化的语言对其进行描述。同时，随着知识的更新，本体的内容也必须作相应的维护^[21]。

2、本体的作用

总的来说，构造本体的目的都是为了实现某种程度的知识共享和重用。本体的作用^[22]主要有以下两方面：

(1)本体的分析澄清了领域知识的结构，从而为知识表示打好基础。本体可以重用，从而避免重复的领域知识分析。

(2)统一的术语和概念使知识共享和重用成为可能。

具体说，本体的作用包括通讯(communication)、互操作(inter-operability)和系统工程(systems engineering)。

(1)通讯：主要为人与人之间或组织与组织之间的通讯提供共同的词汇。

(2)互操作：在不同的建模方法、范式、语言和软件工具之间进行翻译和映射，以实现不同系统之间的互操作和集成。

(3)系统工程：本体分析能够为系统工程提供以下方面的好处：

①重用(re-usability)：本体是领域内重要实体、属性、过程及其相互关系形式化描述的基础。这种形式化描述可成为软件系统中可共享和重用的组件(component)。

②知识获取(knowledge acquisition)：当构造基于知识的系统时，用已有的本体作为起点和基础来指导知识的获取，可以提高其速度和可靠性。

③可靠性(reliability)：形式化的表达使得自动的一致性检查成为可能，从而提高了软件的可靠性。

④规范描述(specification)：本体分析有助于确定IT系统(如知识库)的需求和规范。

3、本体的分类

由于本体的分类方法很多，目前还没有能够被广泛接受的分类标准。但以下几个概念的定义意义明确，并从某种程度上提供了本体的分类方法：

(1)领域本体(DOMAIN ONTOLOGY)：以某一领域为描述对象的本体(区别于领域的问题和任务)。

(2)问题解决模型(PROBLEM SOLVING MODEL)：以问题解决方法为描述对象的本体。

(3)表示本体(REPRESENTATION ONTOLOGY)：以知识表示语言为描述对象的本

体。在表示本体中，类、对象、关系、属性、槽等术语经过严谨的分析和定义。

4、本体的描述语言

本体语言使得用户可以为领域模型编写清晰的、形式化的概念描述，因此它应该满足以下要求：

- ①良好定义的语法 (a well-defined syntax)；
- ②良好定义的语义 (a well-defined semantics)；
- ③有效的推理支持 (efficient reasoning support)；
- ④充分的表达能力 (sufficient expressive power)；
- ⑤表达的方便性 (convenience of expression)。

大量的研究工作者活跃在该领域，因此诞生了许多种本体描述语言^[23]：

①和 Web 相关的有 RDF 和 RDF-S、OIL、DAML、OWL、SHOE、XOL；其中 RDF 和 RDF-S、OIL、DAML、OWL、XOL 之间有着密切的联系，是 W3C 的本体语言栈中的不同层次，也都是基于 XML 的。而 SHOE 是基于 HTML 的，在 HTML 的一个扩展。

②和具体系统相关的（基本只在相关项目中使用的）有：Ontolingua、CycL、Loom。

③KIF 已经是美国国家标准，但是它并没有被广泛应用于互联网，作为一种交换格式更多的应用于企业级。

5、本体的应用

由于本体可用来详细、正确地描述事物的属性，因此本体可以广泛应用于知识工程、数据库设计与集成、信息获取与挖掘等领域。在软件设计、数据库设计领域，本体的运用使系统更加灵活，容易扩展、变化；在知识工程、信息挖掘领域，基于本体的系统能够挖掘出对象以及对象之间的关联，是立体化的挖掘，比之传统意义上的基于关键词匹配的系统，挖掘的信息更加丰富、准确。由此可见，本体论已经成为知识工程领域中的一种有效的理论指导和实践工具，通过本体可以很方便地实现知识的共享和重用^[24]。

1.3.2 偏差分析工作的发展

1、偏差的定义与分类

偏差是对差异和极端特例的描述，揭示事物偏离常规的异常现象。数据库中的数据存在很多偏差，而偏差中又包括很多潜在而有用的知识，如分类中的反常实例、不满足规则的特例、观测结果与模型预测值的偏差、量值随时间的变化等

[26]。

偏差通常可以分为两类：一类是需要丢弃的无用数据，通常由于数据中的错误导致；一类是没有错误、甚至含有重要信息的数据^[26, 27]。对于偏差进行分析在数据分析工作中是很重要而有意义的，原因在于：这些偏差对分析的结果有很大的影响；有一些偏差在应用领域体现了一些有趣的现象或重要的信息；对于大多数应用来说，识别出来的偏差通常能导致意外的知识的发现，而这些知识可以用来检测过去产生的偏差或预测将来的可能趋势、发展方向。因此，发现数据中存在的偏差是非常有必要的。

2、偏差分析的方法

偏差分析(deviation analysis)作为非常重要的数据分析工作之一，就是把数据现状和历史纪录或标准作一个比较，找出它们的偏差，分析偏差的大小及产生偏差的原因。其基本思想是寻找观察结果与参照量之间的有意义的差别。观察结果常常是某一个域的值或多个域值的汇总；参照是给定模型的预测，外界提供的标准或另一个观察。

用于偏差检测的方法很多，其中数据挖掘本身的其它方法也可用于偏差分析，例如关联规则挖掘方法、聚类方法等。但关联规则用于偏差分析的方法还不是很完善，存在许多不足。聚类是用于偏差分析的一种很好的技术。聚类增强了人们对客观现实的认识，是概念描述和偏差分析的先决条件。

目前的偏差分析主要是检测或寻找海量数据中出现的含有重要信息的异常点(outlier)，这在数据挖掘的研究和应用方面已受到广泛关注。偏差在数据挖掘中也可称其为异常或孤立点。异常值是取值远远偏离于一般值的数据，它的存在往往会对数据挖掘的过程产生误导作用，生成不准确，甚至错误的结果^[28, 29]。因此，需要对异常值进行处理。异常探测和分析的具体方法有以下几种：

①基于统计的方法

统计的方法对给定的数据集合假设了一个分布概率模型（如正态分布），然后根据模型采用不一致性检验来确定孤立点。

②基于距离的方法

基于距离的孤立点可以看作是那些没有足够多邻居的对象，邻居是基于距给定对象的距离来定义的。此方法不需要用户拥有任何领域知识，在概念上更加直观。

③基于偏离的方法

此方法不采用统计检验或基于距离的度量值来确定异常对象，它通过检查一组对象的主要特征来确定孤立点，与这个主要特征背离很大的纪录就被认为是一个孤立点。^[30-32]

3、偏差分析的应用

目前, 偏差分析已应用在电子商务异常、破产、信用卡欺诈, 股票交易中性能统计学分析, 入侵检测系统、高效图书馆管理、其它控制系统或软件的鲁棒性检测等领域中^[33, 34]。这些应用可以帮助我们发现其中的异常数据, 从而有助于我们采取进一步的措施。由此可见, 现有的偏差分析只是针对异常数据进行的, 而对于观测结果与模型预测值的偏差的分析进行得比较少。

除了以上应用, 在偏差的深入分析研究方面, Edwin M. Knorr 和 Raymond T. Ng 不仅提出了一种基于距离的偏差检测方法, 还试图进一步应用内涵知识来解释产生偏差的原因^[36, 38]。由此可见, 目前的偏差分析重点只是找到这些偏差, 而对偏差做进一步分析的研究还很少。

我们的研究不同于以往的偏差分析, 我们主要是对偏差进行分类整理、寻找对偏差的解释。因为我们开发人工智能系统的主要目的是为人类服务, 要让人能够了解海量数据的内容, 亦即, 数据变换为人的知识, 以此来帮助和改善人的解决问题能力和决策能力, 那么就需要将偏差有条理地组织起来、寻找产生偏差的原因解释。因此, 我们的研究针对所有与模型预测值有偏差的数据, 尤其是那些含有重要信息的偏差, 不仅仅是找到数据中的异常点, 重点在于还要对偏差进行分类整理、进一步分析偏差的产生原因并对存在正当理由的偏差给出人们可以理解的解释, 从而为以后的工作提供指导。

总之, 我们需要通过对一些实例数据进行分析, 提出针对解释的偏差分析问题解决方法, 从而有助于得到人们易于理解和具有指导意义的解释。而基于本体的方式可以使此方法具有更高的可重用性。在这里, 我们首先要考虑对偏差分析问题建模(模型化)。

1.4 本文章节安排

第一章介绍了本文的研究背景、研究意义、研究内容和具体的工作, 以及相关领域的研究进展和成果。

第二章介绍了进行本文研究所用到的基础理论知识, 主要有知识工程、启发式分类和问题解决方法。

第三章介绍了如何通过数据空间、解释空间、解释标准、匹配标准、分类层次结构、属性值表和多对多关系等概念, 将面向解释的偏差分析问题模型化。

第四章介绍了在第三章所做工作的基础上提出的面向解释的偏差分析问题解决方法, 还介绍了另一位同学所完成的关于可重用的面向解释的偏差分析系统的本体设计和实现工作。

1.5 本章小结

知识工程从诞生至今,近三十年来已有很大发展。目前,知识工程存在的主要问题是知识的共享和重用。由于本体和问题解决方法的实质性作用都是支持重用,因此将二者结合起来进行研究成为知识工程发展的主流方向。

针对互联网以及众多正在运行的信息管理系统,深入研究本体和问题解决方法,并基于二者构建用于知识处理的语义Web服务,从而有效解决知识的共享和重用问题。我们可以利用这个服务做很多工作,比如数据分析、数据校验等。在这里我们关注数据分析中的重要工作之一偏差分析,提出了面向解释的偏差分析。

本篇论文所做的工作就是面向解释的偏差分析工作的前半部分,主要是通过对一些偏差分析的实例进行分析,模型化面向解释的偏差分析问题,并总结出面向解释的偏差分析问题解决方法,得到带有抽象类别的人易理解的偏差解释。面向解释的偏差分析可纳入知识工程中启发式分类的范畴,但是在具体分析方式上,又有着自己的特点。

第2章 理论基础

2.1 知识工程

知识工程(Knowledge Engineering)最初是由费根鲍姆(Feigenbaum)于1977年提出来的。他认为“知识工程是用人工智能的原理和方法,对那些需要专家知识才能解决的应用难题提供解决的手段。恰当地运用专家知识的获取、表达和推理过程的构成与解释,是设计基于知识的系统的重要技术问题”。因此,知识工程是人工智能发展的产物和重要分支^[37]。

知识工程通常被认为是以知识为处理对象,借用工程化的思想,对如何用人工智能的原理、方法、技术来设计、构造和维护知识型系统的一门学科,是人工智能的一个应用分支。由此有人将知识工程叫做应用人工智能,以表示与早期的理论性人工智能有很大的区别。知识工程的目的是在研究知识的基础上,开发智能系统。因此,知识的获取、知识的表示、知识的运用便构成知识工程的三大要素。

知识工程的核心则是专家系统。因此,我们更关注于针对如何在计算机中组织知识、建立高质量知识库,如何使计算机获取与搜索到有用的知识、使用知识来解决问题,来研究通用的问题表达和解决方法。其中,基于规则的专家系统是较为流行的专家系统模型之一。

基于规则的方法是根据以往专家诊断的经验,将其归纳成规则,通过经验知识进行推理。它具有明确的前提,得到确定的结果。它是构建专家系统最常用的方法。早期的专家系统大多数是用规则推理的方法,如DENDRAL专家系统、MYCIN专家系统、PROSPECTOR专家系统等。基于规则的方法容易使知识工程师与人类专家合作,易于被人类专家理解,统一的格式便于管理和推理机的设计。但它也有诸多缺点,如没有描述出要解决的是什么类型的问题,规则间的互相关系不明显,知识的整体形象难以把握、处理效率低、推理缺乏灵活性^[38]。它只适合开发规模较小的系统,对于复杂系统难以用结构化数据来表达,如果全部用规则的形式来表达,不仅提炼规则相当困难,而且规则库将十分庞大和复杂,容易产生“组合爆炸”。它在实时处理方面的应用也已被证明比较困难。

通过对基于规则的方法进行归纳总结,提炼出了启发式分类的思想,从而在更高的抽象层面上描述了一个通用的推理行为。早期知识工程的研究侧重于知识的表示及推理机制。但是目前为了更好地解决问题,知识工程的重点在于实现知识的共享和重用。因此,仅仅将知识进行简单的“堆砌”是不够的,还需要对不同层次的知识进行分类、整理,并有条理地组织起来,使其便于利用、维护和重

用。启发式分类就描述了这样的知识处理的思想。

由于本体和源于启发式分类的问题解决方法的实质性作用都是支持重用,因此将二者结合起来进行研究成为知识工程发展的主流方向、有助于提高人们处理问题和解决问题的能力。

2.2 启发式分类

2.2.1 简单分类

最简单的分类是将一些未知的对象或现象识别为一组已知类别的对象、事件或进程。比较有代表性的是,这些类是分层次组织的典型类型,识别的过程是未知实体的观察值与已知类别特征的匹配过程。一个例子是利用参考手册里诸如颜色、结构和大小等特征识别植物或动物。MYCIN 系统通过将数据信息与细菌层次进行匹配,从而解决了如何从实验数据中识别未知生物体的问题(见图 2-1)。

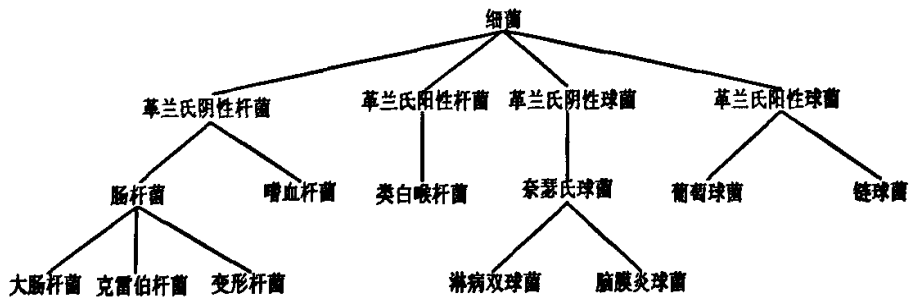


图2-1 Mycin的细菌分类

Figure 2-1 Mycin's classification of bacteria

这个分类的本质特征是问题解决程序从一组事先列举的解决方案中选取。当然,这并不意味着“正确的答案”一定是这些解决方案中的一个,问题解决程序仅仅是试图将数据与已知解决方案进行匹配,而不是构造一个新的解决方案。证据可以是不确定的且部分匹配的,因此输出的可能是一些假设的有序列表。除了匹配之外,还有一些推断解决方案的推理规则。

2.2.2 启发式分类

在上个世纪的70年代末到80年代初,各种启发式程序(通常叫做“专家系统”)被开发用来解决科学、工程、商业和医学等各个不同领域的问题。但是对于描述问题类型和程序设计语言之间映射的尝试并不是很令人满意,因为它们没有描述出一个给定语言知道些什么:类似于“诊断”的面向应用的描述都太笼统了(例

如解决一个诊断问题不一定需要一个设备模型), 而类似于“基于规则”的科技术语并没有描述出要解决的是什么类型的问题。因此, 需要一个关于启发式程序做些什么和知道些什么的更好描述-一个关于它们能力的计算化描述-独立于任务、独立于程序设计语言的实现。逻辑学被认为是“知识水平”分析的基础, 它可以明确启发式程序该做些什么和可能知道些什么。然而, 当时缺乏大量的做这件事的术语和关系。

因此, William J. Clancey提出了启发式分类(Heuristic Classification)的知识处理思想^[4]。所谓启发式分类, 就是通过数据抽象、启发式匹配和细化(refinement)三种操作来对数据进行分类, 其中启发式匹配是指将数据启发式地映射到事先列举的解决方案层次上。与简单分类的不同之处是在不同分类层次上的相关概念是通过非层次、不确定推断得到的。

启发式分类的推理结构如图2-2所示, 其中推理方向、“抽象”与“细化”(refinement)的关系均是简化了的。

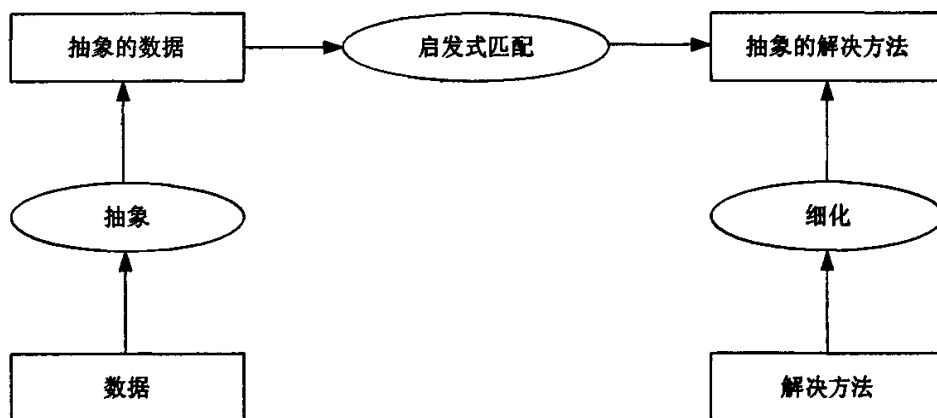


图2-2 启发式分类的推理结构

Figure 2-2 Inference structure of heuristic classification

启发式分类模型描述了关于常见问题和解决方案的一种知识形式和推理模式是启发式相关的。在获得了可能发生的问题和可能产生的解决方案后, 这个知识本质上是根据经验得来的、有着独立于问题领域的总体形式。启发式分类是一个计算方法, 而不是一个要解决的问题。因此, 叫做“启发式分类方法”而不是“分类问题”。启发式分类模型主要通过抽象、匹配和细化(refinement)三种操作来解决问题。

(1) 抽象

在最简单的分类问题中, 数据是具有解决方案特征的, 所以匹配过程是直接的。但是, 对于许多问题, 解决方案的特征并不显示为数据, 而是通过数据抽象推理得到。在启发式程序中抽象数据有以下3个基本关系:

- 定义性抽象：基于概念的本质且必要的特征(“如果结构是一维网络，那么其形状为柱形”);
- 定性的抽象：一种含有定量数据的定义形式，通常与一些标准或期望值有关(“如果病人是成人且白血球数小于 2500，那么白血球数低”);
- 泛化：在子类型层次中泛化(“如果客户是法官，那么他是一个受过教育的人”);

这些解释通常是由确定性程序给出的；还要选择置信区间和限定条件，使得抽象是可区分的。一般都把这种知识看作“真实的”或“确定的”。

数据抽象就是根据领域知识从已有观测值中推理出新的观测值，将一组观测值作为输入、并产生一个观测值作为输出。在数据空间中隐含着一种层次结构：直接观测到的数据是这一层次结构的最底层，高层的数据是通过低层的数据进行抽象得到的。

(2) 细化

解决方案的细化就是以—个解决方案作为输入并产生多个更加具体化的解决方案-即：含有一个现象子集的解决方案可以由“父辈”的解决方案所解释。在这个过程中要应用到关于解决方案的领域知识。细化可以被看作是抽象的逆过程：不是从具体数据到抽象数据，而是从抽象的解决方案到更具体的解决方案。细化引入了一个解决方案空间上的层次结构来表示细化空间。

(3) 启发式匹配

在简单的分类中，数据可以直接匹配解决方案特征或在抽象后匹配。在启发式分类中，解决方案和解决方案特征也要通过另一个分类层次中的一些概念的直接、非层次关联来启发式地匹配。例如，MYCIN 根据生物体的显著特征不仅仅能识别一个未知生物体：MYCIN 还启发式地将病人的抽象描述与疾病类别进行匹配。

一个启发式关系是不确定的、基于典型性假设的、而且有时仅仅是个不易于理解的相互关联。启发式方法通常是根据经验的、由问题解决经验中得来。这种类型的启发式方法通过忽略中间关系(intermediate relations)减少了搜索(这就是为什么没有把抽象关系叫做“启发式法”的原因)。这些关联通常是不确定的，因为这些中间关系可能在特定的例子中不具备。中间关系会被忽略掉是因为它们很难被察觉或很难理解。在医学诊断程序中，启发式方法很典型地跳过了症状和疾病之间的因果关系。

总之，在启发式分类中抽象的数据描述与特定的问题解决方案或描述解决方案的特征是相关联的。启发式分类揭示了一种在推理模式中存在的普遍行为，是问题解决方法的思想根源。

2.3 问题解决方法

2.3.1 基本概念

问题解决方法(Problem Solving Methods, PSMs)^[6]在当今被看作是构建知识库系统(KBS)的主要组件。问题解决方法描述了一个独立于实现和领域(implementation- and domain-independent)的知识库系统(KBS)的推理过程,其思想最早源于启发式分类。一个问题解决方法的描述如下:

- 问题解决方法明确了为了实现任务的目标需要执行什么样的推理行为。
- 问题解决方法定义了在这些行为上的一个或多个控制结构。
- 知识角色明确了领域知识在每个推理行为中所扮演的角色。这些知识角色定义了独立于领域的通用术语。主要有两类角色:静态角色-描述了问题解决方法所需的领域知识;动态角色-形成了推理行为的输入和输出。

问题解决方法有以下几个目的:

- ①知识库系统构建(知识工程):一个问题解决方法对于构建用于描述实现特殊任务目标的问题解决程序是非常有用的。通常,这里含有一个任务分解方法。
- ②知识库系统规范(推理):一个问题解决方法可以描述一个用于实现任务目标的有效推理过程。在这个意义上,一个问题解决方法关心的是构建过程的产物,这个方法与知识库系统的设计模型相关。
- ③认知建模:一个问题解决方法可以描述人类问题解决的认知模型。

问题解决方法在知识工程和知识获取中扮演着重要角色。它可以通过应用领域知识来有效地实现目标,引导领域知识的获取过程,并通过重复使用来推动知识库系统(KBS)的开发。用于开发问题解决方法的的不同方法包括:基于人类问题求解的分析、基于效率关注的分析,基于可用领域知识的分析。

问题解决方法的重用通常包括以下问题:存在什么样的通用问题解决方法,以及这些方法是如何组织的?为了支持所选择的特定应用,问题解决方法是如何以一种方式索引的?我们如何让通用问题解决方法适用于特定应用环境?如何将知识库中的单个问题解决方法配置成一个一致的问题求解程序?

2.3.2 问题解决方法的体系结构

问题解决方法一般由三个相关部分组成,描述了问题解决方法能实现什么目标、它是如何实现这个目标的、为了实现这个目标需要些什么,这分别表示为问题解决方法的能力(competence)、操作规范(operational specification)和需求/假

设 (requirements/assumptions) (见图 2-3)。

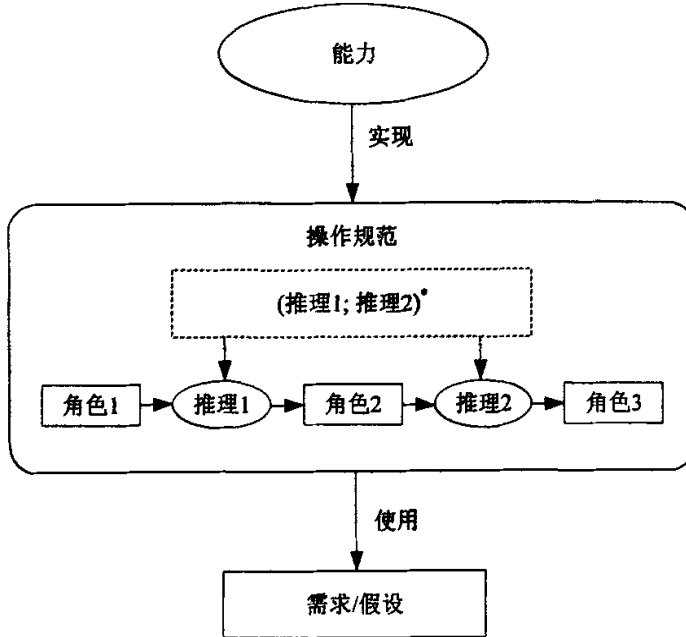


Figure 2-3 The architecture of a PSM

- 能力 (Competence)。一个问题解决方法的能力是输入输出行为的说明性描述，描述了此方法能实现什么。
- 需求/假设 (Requirements/assumptions)。一个问题解决方法的需求/假设描述了为了实现其能力，此方法所需的领域知识。例如在参数设计任务中的需求包括：连接违规约束和可能修复行为的启发式的可用性，一个优先选择关系必须描述一个完整排序的事实。这些需求描述了为了实现能力，问题解决方法需要提供些什么。
- 操作规范 (Operational specification)。一个问题解决方法的操作规范描述了在提供了所需知识的情况下，传递特定能力的推理过程。它由推理步骤、知识和控制流程组成。推理步骤明确了所有实现此方法能力的推理步骤。这些步骤通过它们的输入/输出关系来描述，可以通过一个方法（也就是说问题解决方法可以层次地分解）或一个基本推理（即不能进一步分解的原子推理步骤）得到。知识的流动通过动态角色得以发生，用来存储输入和输出推理的行为。最后，一个问题解决方法的控制描述了推理步骤的执行顺序。

2.3.3 问题解决方法与本体的结合

问题解决方法描述了一个通用的推理行为，这一推理行为可以重用在不同的

应用领域上。因为知识库系统是由领域知识和问题解决知识共同组成的，本体在一般层面上定义了静态领域知识，而问题解决方法指定了一般动态推理知识，因此，有必要把本体和问题解决方法放在一起研究、并关注于二者的集成。二者的结合可以推动这两个领域的发展，有助于知识库系统中知识的共享和重用，推动知识库系统的实际应用。

由于本体和问题解决方法的实质性作用都是支持重用，所以基于本体的问题解决方法及其相关的方法论研究一直引人关注：

1、OntoKADS

Sabine Bruaux等人提出了一个以本体为中心的用于构建专家经验模型的方法-OntoKADS。这个方法是建立在一个区分了两个概念化层面的核心问题解决本体上的：在“对象层面”上，OntoKADS提供了一组定义问题解决情况类的概念。在“元层面”上，OntoKADS提供了一组用于对CommonKADS建模原语进行编码及定义相应专家经验模型的元概念。OntoKADS方法在很大程度上把专家经验模型的构建看作本体的构建。此方法由两步组成：第一步，开发了通过建模原语来描述概念的问题解决驱动应用本体；第二步，进一步明确了与任务相连接的问题解决方法^[10, 11]。

2、UPML

UPML（统一问题解决方法开发语言）是问题解决方法和本体的方法集成。它给出了一个全面的基于通用问题解决方法的框架。根据这个框架，每个问题解决方法都可以根据三个独立的维来描述：领域维、任务维和问题解决范例维。领域维包括应用问题解决方法所需的知识结构的类型和知识属性，还表示了领域是独立于还是依赖于问题解决方法。任务维表示了问题解决方法对于特定任务的依赖关系。问题解决维包括问题解决方法的算法结构^[39]。

3、基于本体的分类问题解决方法

因为启发式分类是问题解决方法思想的根源，所以由启发式分类引出的关于分类问题的解决方法也受到广泛的关注^[4, 40, 41]，这在许多文献中已有了详细的分析^[42, 43]。其中，Enrico Motta等人也以启发式分类为基础来对分类问题进行建模、并提出了基于本体的分类问题解决方法^[44]。分类可以看作是根据一些标准来寻找解决方案（类）的问题，这一解决方案（类）能够最佳地解释关于未知对象的一组特定的已知事实（观测值）。分类问题可以通过以下概念来模型化：观测值、解决方案、匹配标准、解决方案标准和分数比较标准。分类问题解决可以被看作是一个根据给定的应用标准来寻找解决方案空间的过程。他们以启发式分类模型作为分类问题解决方法的基础。更简单的问题解决方法可以定义为启发式分类模

型的“退化”-例如：这些问题解决方法可能没有包括抽象和（或）细化机制。更复杂的问题解决方法是在这个模型上添加额外的推理机制。分类问题解决的方法本体包括抽象器、细化器、分类任务的本体（包括观测值的定义、解决方案的定义和匹配标准的定义，还有用于这些概念的相应关系和函数）。其中，抽象器是根据领域知识从已有观测值中推理出新的观测值，细化器是以一个解决方案作为输入并产生多个更加具体化的解决方案。2003年，Motta等人基于UPML又完成了互联网推理服务系统（IRS）^[46]，使用此系统可在线配置问题求解资源。随后的工作直到2007年，一直是在研制建立因特网上自助或半自动的语义服务系统。

面向解释的偏差分析给出了一个基于本体的偏差分析问题解决方法，通过计算观测值与标准值之间的偏差值并对其进行分析来寻找产生偏差的原因。实际上，系统得到的数据存在于一个数据空间，而产生偏差的解释在一个解释空间，寻找解释就要在这两个空间中建立联系。这种工作正是符合由William J. Clancey提出的启发式分类模型的。因此，面向解释的偏差分析可以纳入知识工程中启发式分类的范畴，但是在匹配的方式、抽象与细化的方法上，仍有自己的特点。与Enrico Motta等人以启发式分类为基础将分类问题模型化相比，我们针对偏差分析问题进行模型化、并给出解决方法：用不同的抽象方法给出了较为规范的抽象过程；采用属性和值分开的方法，通过应用解释标准在解释空间进行匹配；使用匹配标准在由解释属性构造的分类层次结构上搜索类别并在属性值表上区分影响程度，完成细化(refinement)过程，最终得到带有抽象类别的人易理解的偏差解释。

2.4 本章小结

知识工程是人工智能发展的产物和重要分支，目前其重点是知识共享和重用。因此需要研究通用的问题表达和解决方法。但是传统的基于规则的系统对知识的整体形象难以把握、处理效率低。因此，通过对基于规则的方法进行归纳总结，提炼出了启发式分类的知识处理思想，来对不同层次的知识进行分类、整理、组织，使其便于利用、维护和重用。

启发式分类，就是通过数据抽象、启发式匹配和细化（refinement）三种操作来对数据进行分类。其中启发式匹配是指将数据启发式地映射到事先列举的解决方案层次上，通过忽略掉一些很难被察觉和理解的中间关系来减少搜索。启发式分类揭示了一种在推理模式中存在的普遍行为，是问题解决方法的思想根源。

问题解决方法描述了一个独立于实现和领域的知识库系统(KBS)的通用推理行为，它可以重用在不同的应用领域上。由于本体和问题解决方法的实质性作用都是支持重用，二者的结合有助于知识库系统中知识的共享和重用。因此基于

本体的问题解决方法及其相关的方法论研究一直引人关注。其中，由启发式分类引出的关于分类问题的模型化和解决方法最具有代表性。

面向解释的偏差分析给出了一个基于本体的偏差分析问题解决方法，可以纳入启发式分类的范畴。与 Enrico Motta 等人以启发式分类为基础将分类问题模型化相比，我们针对偏差分析问题进行模型化、给出解决方法，最终得到带有抽象类别的人易理解的偏差解释，其中在匹配的方式、抽象与细化的方法上具有自己的特点。

第3章 面向解释的偏差分析的模型化

3.1 优势、原理和缺陷

知识工程中的启发式分类方法揭示了一种在推理模式中存在的普遍行为,是问题解决方法的思想根源。它对于以下方面的实践具有重要的意义:构造专家系统,设计新的工具和在此领域的不断研究。但是它是从知识密集型领域抽象出来的方法,其抽象与细化过程都需要较多的领域知识,而且这些领域知识还要具有合适的层次结构,因此具有较大的难度并需要较高的花费。这使得此方法在应用时具有一定的局限性。

为了使启发式分类方法能应用于更多的领域,同时为了能快速有效地帮助和改善人的解决问题能力和决策能力,我们提出了面向解释的偏差分析,这是启发式分类方法的一种。这种偏差分析的优势在于:其所涉及的一些领域中,问题通常并不很复杂,但一般没有提供具有适宜分类层次的领域知识;尤其适用于一些对处理时间有一定限制的领域;而且是一些实际数据与预估结果有所偏差而需要解释的领域。在此方法中,我们主要针对三个这样的领域来进行分析:试卷成绩分析领域、超市经营数据分析领域、球赛统计数据领域,其中球赛统计数据领域对处理时间有较高的要求。由于我们的方法不需要太多的领域知识、有较好的时效性,所以更加实用、处理问题更加方便快捷,在一定程度上扩展了启发式分类方法的应用范围。而且此方法在后续工作中与本体的结合,使得我们的方法具有较高的可重用性和更普遍的应用价值。

面向解释的偏差分析的原理就是通过计算得到观测值与标准值之间的偏差值,并对其进行分析来寻找产生偏差的原因。实际上,系统得到的数据存在于一个数据空间,而产生偏差的解释则在一个解释空间,寻找解释就要在这两个空间中建立联系。这种工作正是符合由William J. Clancey提出的启发式分类模型的。启发式分类方法和面向解释的偏差分析方法都有一个启发式的匹配过程,因此我们将后者纳入前者的范畴、面向解释的偏差分析工作应适合在知识工程领域中加以探讨。但是在匹配的方式、抽象与细化的方法上,后者仍有自己的特点。这也是本文将要论述的重点。

整个面向解释的偏差分析过程就是针对一些应用领域没有提供具有适宜分类层次的领域知识的情况,采用属性和值分开的方法加以解决,主要分为三步:首先,输入原始数据,在数据空间中采用不同于经典启发式分类的一些抽象方法抽象出偏差值;其次,根据数据空间中的偏差值,通过启发式推断,得到一个由一组解释值构成的解释;然后,使用解释标准和匹配标准在解释空间中由解释属

性构造的分类层次结构上搜索这个解释的类别并在属性值表上区分这个解释的影响程度，完成细化过程。最终得到带有抽象类别的人易理解的偏差解释。

当然，面向解释的偏差分析也有一些不尽人意的地方，其中有些发生可能性较小的或是不合理的偏差情况是被我们忽略的，有待进一步研究。我们只针对试卷成绩分析、超市经营数据分析、球赛统计数据领域进行了研究，还需要加强对更多实际领域的研究、扩展可重用领域，从而获得更易于人们理解的解释、使该方法更具有普遍性。因此，在偏差分析这个领域上，我们还需要更深入、更广泛的研究，从而更准确、更高效地解决问题。

3.2 模型化面向解释的偏差分析问题

面向解释的偏差分析的模型化是面向解释的偏差分析研究的前半部分工作，是对偏差分析问题的形式化描述。它是提出问题解决方法的前提，也是后半部分工作的基础、构造本体的依据和前期准备工作。

接下来，我们将主要通过数据空间、解释空间、解释标准、匹配标准、分类层次结构、属性值表和多对多关系等概念来描述偏差分析问题。

3.2.1 数据空间

我们将有待分析的偏差存放于一个数据空间中，而偏差又是根据数据空间中的数据得来的。因此，在数据空间，涉及了三个基本概念：观测值、标准值和偏差值。它们都可以描述为 $\langle f, v \rangle$ 的形式，其中 f 是一个未知对象的属性， v 是这一属性的值。本文称它们为一个属性值对，并采用一般的观点来看待属性这个概念：属性可以描述一个对象，其值是可以直接观察到或可以推出来的。

1、观测值

是主要的输入数据。根据得到的方式是直接还是间接的，我们可以将观测值分为基础数据值和统计数据值。基础数据值是实际直接观测到的，即指对一些事物或现象测量所得的原始数据；统计数据值是根据基础数据进行整理得到的。通常，统计数据具有以下特点：

①统计数据具有“二合一”性

统计数据是概念名和数值的统一体，是不可分割、缺一不可的。

②统计数据具有历史性

统计数据是历史发展的积累，随着时间的推移，以往的历史数据不会失去存在的意义，而是进行统计分析、统计预测的基础。

③统计数据具有广泛性

统计数据所记录的对象可能横向涉及到各行业的各种事物,而且随着统计手段的加强和统计能力的提高,统计的范围在不断拓展。

④统计数据具有大量性

统计数据的纵向历史性和横向广泛性造成统计数据的大量性。

⑤统计数据具有多维性和层次性

分析单个统计数据,可以看出它的概念名具有结构多维性。即:一个概念名是由多个基本元素构成的。分析多个统计数据,概念名之间又具有结构层次性。

2、标准值

可以是长期观测的平均值,也可以是根据这些平均值在实际需求下加以修改的预测值,还可以是在无历史纪录时通过对一些属性进行分析给出的预测估计值。第一种情况是可以直接获得的,后两种情况是间接获得的。标准值是一种参照量。

3、偏差值

是将观测值与标准值进行比较,若得到的偏差数在一个允许的范围内,偏差值为空,即没有偏差;否则,偏差值可取大小、高低等值来表示偏差。一般出现偏差的属性越少,实际情况与标准或预测越接近,如无一偏差,则完全吻合,无需解释。如果存在偏差,就需要对它进行解释。因为偏差中通常包括很多潜在而有用的知识,对它的解释可以指导我们的工作。

启发式分类模型的抽象操作一般可通过规则的形式来表达,主要有三个基本方法:定义性抽象、定性抽象和泛化抽象。在数据空间中,面向解释的偏差分析是从观测值和标准值的比较中得到偏差值,相当于做了一次抽象。也就是说,我们的偏差分析将一般启发式分类的多次数据抽象规范成一次抽象。但对于某些领域,由于观测值与标准值不能直接得到,此时数据空间中的抽象会多于一次。比如,对基础数据进行统计得到所需的观测值就是一次或多次抽象,对历史纪录进行修改得到或对一些数据进行分析后预测得到所需的标准值也是一次或多次抽象。所以,在数据空间中,得到偏差值需要做的工作如图3-1所示:

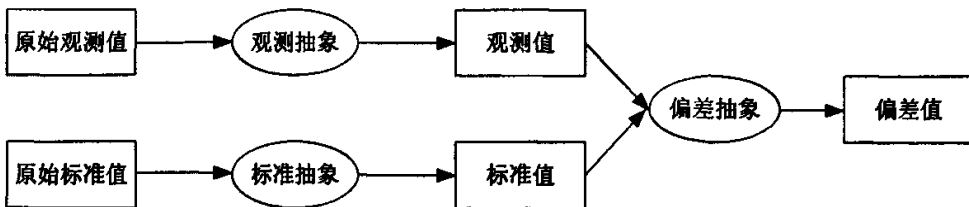


图3-1 数据空间中的抽象

Figure 3-1 The abstraction in Data Space

接下来，具体针对所研究的三个领域来分析这种情况：

1、观测值的抽象

(1) 针对试卷成绩分析系统

①原始观测值

从试卷中直接获得的学生实际成绩，比如每个学生的学号、姓名、每道题的得分和整个试卷的得分。

②观测值

由教育测量学可知学生的成绩一般呈正态分布，可通过对学生的实际成绩进行统计得到分布状况，此时的分布特性是所需的观测值；此外，实际平均分也是所需的观测值，它是根据实际学生成绩计算出来的。因此，要得到这些所需观测值而进行的抽象可称之为利用已有规律的数学公式的抽象。

其实，对于数据分布特性的测量和描述，可以从三方面进行^[46]：

➤ 分布的集中趋势

反映各个数据向其中心值靠拢或聚集的程度，寻找数据分布的中心值或一般水平的代表值。主要指标有：针对分类数据，可以用众数；针对顺序数据，可以用中位数和分位数；针对数值数据，可以用均值（如简单均值、加权均值、调和平均数和几何平均数等）。

➤ 分布的离散程度

反映各个数据远离其中心值的趋势或差异程度。主要指标有：针对分类数据，可以用异众比率；针对顺序数据，可以用四分位差；针对数值数据，可以用极差、平均差、方差和标准差；针对相对位置的测量，可以用标准分数；针对相对差异程度，可以用离散系数（也叫变异系数）。

➤ 分布的形状

反映数据分布是否对称、集中趋势高低、偏斜程度和扁平程度。主要指标有：偏度和峰度。前者描述数据偏斜的方向和程度；后者描述数据分布平峰或尖峰的程度。

在我们的研究中，主要是从第三方面数据的分布形状来考虑的，所以所需的观测值是偏度和峰度。计算偏度和峰度的公式为^[47]：

$$\text{偏度: } \alpha = \frac{\sum_{i=1}^N (X_i - \bar{X})^3}{N\sigma^3} \quad (3-1)$$

$$\text{峰度: } \beta = \frac{\sum_{i=1}^N (X_i - \bar{X})^4}{N\sigma^4} \quad (3-2)$$

其中, X_i 代表每个学生成绩的实际分数, \bar{X} 代表所有学生成绩的实际算术平均分, σ 是标准差, N 是学生的个数。

从公式中可以看出, 需要用到学生成绩的实际平均分, 实际平均分又是根据实际学生成绩计算出来的。由此可以看出, 要得到所需的观测值(实际平均分、偏度和峰度), 需要进行一次、甚至多次抽象。

(2) 针对超市经营数据分析系统

①原始观测值

超市当年经营状况的实际数据, 比如每天的客流量、销售量、消耗和利润等。

②观测值

通过对原始观测值进行统计, 得到所需的超市当年经营中五个属性(客单价、客单数、平均毛利率、经营费用、利润)的观测值, 其中利润观测值是根据前四个属性的值、通过经济学公式(利润=客单价 x 客单数 x 平均毛利率-经营费用)^[49]计算得到的。这种抽象是一种统计的抽象。其中:

客单价: 是指顾客一次性购买商品的平均销售额。

客单数: 是指有效的客流数, 即来超市后消费的客流数。

平均毛利率: 是指所有商品销售毛利的平均值, 平均毛利率=毛利额/销售额。

经营费用: 包括可控的经营费用和不可控的经营费用。

(3) 针对球赛统计数据系统

①原始观测值

是足球比赛中所记录的实际数据, 例如比赛中的得分情况、裁判的执法情况、传球和射门的情况等。

②观测值

通过对实际比赛中所记录的数据进行统计, 得到所需的观测值: 比赛分数、判罚次数、控球时间、我方射门次数、对方射门次数和射正率。这种抽象是一种统计的抽象。其中:

比赛分数: 实际比赛中总的得分是多少。

判罚次数: 实际比赛中被罚的次数。

控球时间：是指在比赛中一方队员拥有球的时间，也就是带球的时间。

我方射门次数：我方球队在比赛中总共射了多少次门。

对方射门次数：对方球队在比赛中总共射了多少次门。

射正率：与前锋的实力或对方的防守相关。射正率=射中次数/射门次数。

2、标准值的抽象

(1) 针对试卷成绩分析系统

①原始标准值

试卷（试题）的满分分数。试卷成绩所呈现的一般正态分布规律。

②标准值

在原始标准值的基础上，通过综合考虑试卷（试题）的本身难度、教学难度和教学要求，来对试卷（试题）的平均分进行预测。此时获得的预测平均分是所需的标准值，可看成是一种统计的抽象。针对分布特性来说，标准偏度和标准峰度是所需的标准值，这是一种对一般规律的抽象。

(2) 针对超市经营数据分析系统

①原始标准值

超市上一年经营中四个属性（客单价、客单数、平均毛利率、经营费用）的实际值。

②标准值

将原始标准值与相关参数进行计算，就可以计算出该超市当年经营中四个属性的预测值，也就是所需的标准值。这些参数主要包括：累计社会消费品零售额增长率，人口增长率，人均收入增长率。这些参数对四个属性的影响：

客单价当年预测值=客单价上一年实际值 \times (1+累计社会消费品零售额增长率)

客单数当年预测值=客单数上一年实际值 \times (1+人口增长率)

经营费用当年标准值=经营费用上一年实际值 \times (1+人均收入增长率)

通过以上三个公式，就能计算出所需的客单价预测值，客单数预测值和经营费用预测值。而毛利率的预测值与上一年毛利率的实际值是基本持平的。之后，就可以通过超市当年经营中四个属性的预测值，用经济学公式（利润=客单价 \times 客单数 \times 平均毛利率-经营费用）^[48]计算出当年中利润的预测值。对预测值（标准值）的抽象可以看成是利用已有规律的经济学公式的抽象。

(3) 针对球赛统计数据系统

因为在球赛中所涉及到的标准值是长期以来比赛中的一个平均值、一种约定或规范，所以没有原始标准值一说，直接就是所需的标准值，主要包括：

比赛分数、判罚次数、控球时间、我方射门次数、对方射门次数和射正率。其中：

- ▶ 比赛分数的标准值：以对方球队在实际比赛中总的得分为标准，与这个标准上下相差5分以内的值都算是标准值。
- ▶ 判罚次数的标准值：通常是指裁判对比赛中某一方队员的平均判罚次数，一般为8-10次。
- ▶ 控球时间的标准值：通常为整场比赛时间的一半，即45分钟，与这个标准前后相差10分钟以内的值都算标准值。
- ▶ 我方射门次数的标准值：以对方的射门次数为标准，与此标准上下相差2次以内的值都算标准值。
- ▶ 对方射门次数的标准值：以我方的射门次数为标准，与此标准上下相差2次以内的值都算标准值。
- ▶ 射正率的标准值：一般规定为55%-65%。

3、偏差值的抽象

(1) 针对试卷成绩分析系统

通过将实际分布特性、实际平均分（观测值）和标准分布特性、预测平均分（标准值）进行比较，我们抽象得到一组偏差值：偏度、峰度、实际平均分和预测平均分的关系。

①偏度的取值范围

- 标准偏度：偏度为零，分布对称。也就是说偏差值为空，没有偏差。
- 正偏度：偏度大于零，分布曲线向右延伸拖长尾巴。
- 负偏度：偏度小于零，分布曲线向左延伸拖长尾巴。

②峰度的取值范围

- 标准峰度：峰度等于3，分布曲线为正态峰度。此时，偏差值为空。
- 高狭峰：偏度大于3，分布趋向为尖顶峰度。
- 低阔峰：偏度小于3，分布趋向为平顶峰度。
- U型峰：偏度小于1.8，分布曲线呈U型分布。
- 直线峰：偏度接近1.8时，分布曲线趋于一条水平线。

③实际平均分和预测平均分的关系

- 等于：实际平均分和预测平均分相等、一样。此时，偏差值为空。
- 小于：实际平均分比预测平均分小。
- 大于：实际平均分比预测平均分大。

由于各个偏差值的取值不同，所以一组偏差值的组合情况可能有很多

种。这些组合中有些情况是不可能出现或不合理的、不合法的。因此我们在得到这些偏差值组合后，首先可以根据偏度和峰度的可能取值，去掉一些不可能出现的偏差分布情况；其次根据实际平均分和预测平均分的关系，最终得到那些可能出现的偏差组合情况。

(2) 针对超市经营数据分析系统

将超市当年经营中五个属性的观测值与预测值（标准值）相比较，得出一组偏差值：客单价、客单数、平均毛利率、经营费用、利润。

①客单价的取值范围

- 大：客单价的实际值比预测值大。
- 相等：客单价的实际值与预测值持平。此时，偏差值为空。
- 小：客单价的实际值比预测值小。

②客单数的取值范围

- 大：客单数的实际值比预测值大。
- 相等：客单数的实际值与预测值持平。此时，偏差值为空。
- 小：客单数的实际值比预测值小。

③平均毛利率的取值范围

- 大：平均毛利率的实际值比预测值大。
- 相等：平均毛利率的实际值与预测值持平。此时，偏差值为空。
- 小：平均毛利率的实际值比预测值小。

④经营费用的取值范围

- 大：经营费用的实际值比预测值大。
- 相等：经营费用的实际值与预测值持平。此时，偏差值为空。
- 小：经营费用的实际值比预测值小。

⑤利润的取值范围

- 大：利润的实际值比预测值大。
- 相等：利润的实际值与预测值持平。此时，偏差值为空。
- 小：利润的实际值比预测值小。

由于各个偏差值的取值不同，所以一组偏差值的组合情况可能有很多种。这些组合中有些情况是不可能出现或不合理的、不合法的。在这里我们需要以经济学公式（利润=客单价×客单数×平均毛利率-经营费用）^[48]为标准来验证这组偏差值的合法性。将一组偏差值中的客单价、客单数、平均毛利率、经营费用这四项属性所对应的值代入公式进行运算。再将运算得出的利润值与该组偏差值中所列的利润值相比较。若两个值相等，则表明该组偏差值合法；反之，该组偏差值不合法。通过去除不合法的偏差值组合情况，

我们最终得到所需的合法的偏差值组合。

(3) 针对球赛统计分析系统

通过将观测值和标准值进行比较，我们得到一组偏差值：比赛分数、判罚次数、控球时间、我方射门次数、对方射门次数和射正率。

① 比赛分数的取值范围

- 大：我方的比赛分数比标准值高。
- 相等：我方的比赛分数与标准值相等。此时，偏差值为空。
- 小：我方的比赛分数比标准值低。

② 判罚次数的取值范围

- 大：实际判罚的次数比标准值多。
- 相等：实际判罚的次数与标准值相同。此时，偏差值为空。
- 小：实际判罚的次数比标准值少。

③ 控球时间的取值范围

- 大：实际控球时间大于标准值。
- 相等：实际控球时间等于标准值。此时，偏差值为空。
- 小：实际控球时间小于标准值。

④ 我方射门次数的取值范围

- 大：我方的射门次数比标准值多。
- 相等：我方的射门次数与标准值相等。此时，偏差值为空。
- 小：我方的射门次数比标准值少。

⑤ 对方射门次数的取值范围

- 大：对方的射门次数比标准值多。
- 相等：对方的射门次数与标准值相等。此时，偏差值为空。
- 小：对方的射门次数比标准值少。

⑥ 射正率的取值范围

- 大：实际射正率大于标准值。
- 相等：实际射正率等于标准值。此时，偏差值为空。
- 小：实际射正率小于标准值。

由于各个偏差值的取值不同，所以一组偏差值的组合情况可能有很多种。这些组合中有些情况是不可能出现或不合理的、不合法的。因此，我们在得到这些偏差值组合后，可以通过属性间的约束关系来去除不合理的偏差值组合情况。所谓约束关系是指属性间的相互作用，也就是说在现实中各个属性之间是相互联系、共同起作用的。比如我方射门次数和对方射门次数是互相以对方为标准的，所以这种约束关系造成了二者不可能产生相同的偏差

值；如果在一个偏差组合中这二者产生了相同的偏差值，那么这个偏差值组合就是不合理的、需要被去除。通过去除不合理的偏差值组合情况，我们最终得到所需的合理的偏差值组合。

3.2.2 解释空间

数据空间中存放的是偏差值，而解释空间存放的是产生偏差值的解释。根据数据空间产生的不同属性的偏差值组合，去寻找不同的解释原因。这实际上是将两个空间中的相关概念通过由经验得来的启发式推断建立起联系。这种启发式推断可以用规则来表达，本文称之为对应关系。与数据空间对应的是解释空间，这个空间的属性值对也可描述成 $\langle g, v' \rangle$ 的形式，并称之为解释值，其中 g 是解释中的一个属性， v' 是这个属性的值。多个不同的解释值的组合描述了一个解释。由于面向解释的偏差分析所关心的领域，一般没有提供具有适宜分类层次的领域知识，因此，本文采用属性和值分开的方法，将解释空间描述为通过属性构造的分类层次结构和由属性值构造的属性值表。

1、分类层次结构

在分类层次结构中，一般可以将解释属性划分为若干基本类，从而使得每个基本类中包含若干解释属性。在一般情况下，分类层次结构的构造是以这些基本类作为最底层，这些基本类的两两组合构成上一层，再由三个基本类一组的组合构成再上一层，接着由四个基本类一组的组合继续往上构造，直到最上一层包括所有基本类。如果有 N 个基本类的话，就可以构造 N 层的分类层次结构。

在我们的研究实例中，解释属性可分为三种基本类：

(1) 外在的解释属性（即外在原因）

例如，超市经营数据分析领域所涉及到的周边人民生活水平、周边人口等属性。

(2) 自身的解释属性（即自身原因）

可细分为两类，一种是与精神有关、与战略有关或与抽象层面有关的称为隐性的解释属性；另一种是与物质有关、与战术有关或与具体层面有关的称为显性的解释属性。

①隐性的解释属性：如球赛统计分析领域涉及到的运动员心态。

②显性的解释属性：如超市经营数据分析领域涉及到的商品进价和商品售价等。

由此可以看出，由可划分为三种基本类的解释属性所构成的解释属性集合共分为8类。比如将外在的解释属性记为 a ，隐性的自身解释属性记为 b ，显性的自

身解释属性记为c，那么解释属性的8种组合情况如表3-1所示：

表3-1 解释的分类

Table 3-1 The classification of explanation

类别	解释属性的组合
第1类	{ \emptyset }
第2类	{a}
第3类	{b}
第4类	{c}
第5类	{ab}
第6类	{ac}
第7类	{bc}
第8类	{abc}

其中{ \emptyset }代表数据空间中各描述属性值均无偏差的情况(即偏差值均为空)，此类无需解释，在分类层次结构中不出现。只有存在偏差的其余7种解释属性组合构成了我们实例中的分类层次结构：(如图3-2)

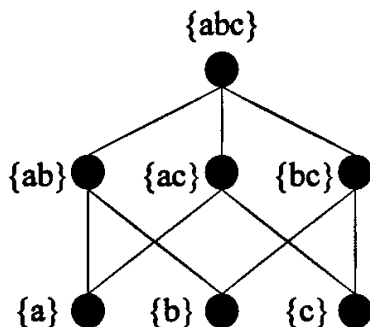


图3-2 分类层次结构

Figure 3-2 The hierarchy structure for classification

2、属性值表

根据面向解释的偏差分析的特点，在给定具体的标准后，产生的偏差有两种情况：正偏差和负偏差：

①正偏差：一般是指对所求问题产生好的或正面的影响的偏差；

②负偏差：一般是指对所求问题产生坏的或负面的影响的偏差。

对应到解释上，正偏差产生正解释，负偏差产生负解释。采用一些正解释的解释值构成的表称之为属性值表。

启发式分类模型在匹配之后，若有具有分类层次的领域知识的支持，则有一

个细化的过程，这是一个抽象的反过程，即由抽象到具体。而面向解释的偏差分析在匹配过程中，将会在解释空间中的以属性构建的分类层次结构上进行搜索，找到解释的类别，然后使用与偏差值对应的解释值，在属性值表上匹配，代替细化的过程。

接下来，具体针对所研究的三个领域进行分析：

1、针对试卷成绩分析系统

(1) 解释属性

在解释空间中的属性主要有：

- ①教学效果：取值范围是好、差。
- ②试题难度：取值范围是高、低。
- ③试题梯度：取值范围是大、小。
- ④学生素质：取值范围是高、低。
- ⑤学生能力：取值范围是强、弱。
- ⑥学生素质分布：取值范围是分散、接近、平均、两极分化。
- ⑦学生能力分布：取值范围是分散、接近、平均、两极分化。

(2) 对应关系

要得到数据空间中一组偏差值对应到解释空间中的解释（多个偏差值解释的组合），首先要对每个偏差值进行分析，所用到的对应关系（规则）是：

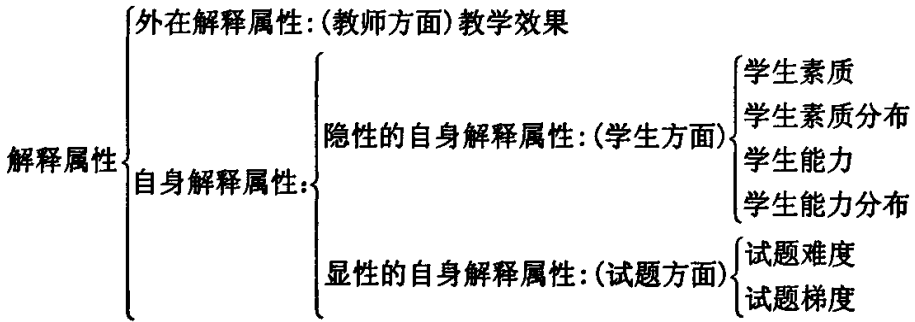
- ①高狭峰—试题梯度小∨学生能力分布接近∨学生素质分布接近
- ②低阔峰—试题梯度大∨学生能力分布分散∨学生素质分布分散
- ③U型峰—学生能力分布呈两极分化差别大∨学生素质分布呈两极分化差别大
- ④直线峰—学生能力分布平均∨学生素质分布平均
- ⑤正偏度—试题难度高∨教学效果差∨学生能力弱∨学生素质低
- ⑥负偏度—试题难度低∨教学效果好∨学生能力强∨学生素质高

然后将单独对每个偏差值分析得到的解释组合起来形成这组偏差值的一个整体解释，而所有的偏差值组合及其对应的解释形成了对应规则库。

通过试卷成绩分析系统的对应规则库，我们可以找到这组偏差值的一个解释（由一组解释值构成）。然后，我们对这个解释进行多对多关系的处理（见第3.2.5部分），达到简化解释的目的。

(3) 解释属性的分类

在得到一组简化的解释值后，我们通过分析可以把这些解释属性划分为三个基本类中的一种：



2、针对超市经营数据分析系统

(1) 解释属性

在解释空间中的属性主要有:

- ①周边人民生活水平: 超市周边人民生活质量的水平。主要表现在人均消费水平上。取值范围是: 提高、降低。
- ②周边人口: 超市周边常住人口数。取值范围是: 增多、减少。
- ③超市员工数: 超市员工的增员或裁员。与超市客流量密切相关。取值范围是: 增多、减少。
- ④超市员工奖金: 超市员工工资中与超市经营状况密切相关的部分。取值范围是: 增加、减少。
- ⑤经营必须消耗: 超市经营中消耗的必需资金。包括水电费, 租金等。取值范围是: 增多、减少。
- ⑥商品进价: 商品的进货价格。取值范围是: 升高、降低。
- ⑦商品售价: 商品的销售价格。取值范围是: 升高、降低。

(2) 对应关系

要得到数据空间中一组偏差值对应到解释空间中的解释(多个偏差值解释的组合), 首先要对每个偏差值进行分析, 所用到的对应关系(规则)是:

- ①客单价大—人民生活水平提高∨商品售价降低
- ②客单价小—人民生活水平降低∨商品售价升高
- ③客单数大—人口增多∨商品售价降低
- ④客单数小—人口减少∨商品售价升高
- ⑤平均毛利率大—商品进价降低∨商品售价升高
- ⑥平均毛利率小—商品进价升高∨商品售价降低
- ⑦经营费用大—经营必须消耗增多∨员工增多∨员工奖金增加
- ⑧经营费用小—员工奖金减少∨员工减少∨经营必须消耗减少

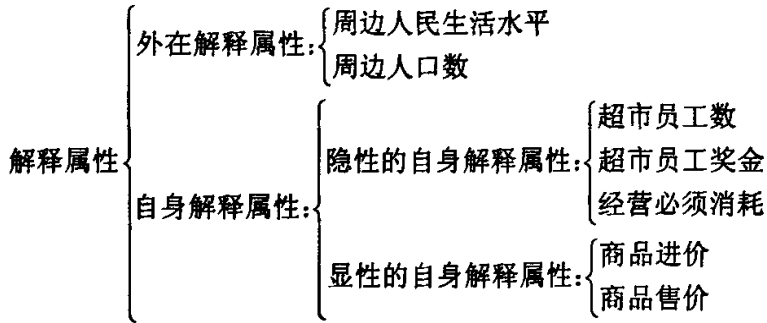
然后将单独对每个偏差值分析得到的解释组合起来形成这组偏差值的一

个整体解释，而所有的偏差值组合及其对应的解释形成了对应规则库。

通过超市经营数据分析系统的对应规则库，我们可以找到这组偏差值的一个解释（由一组解释值构成）。然后，我们对这个解释进行多对多关系的处理（见第 3.2.5 部分），达到简化解释的目的。

(3) 解释属性的分类

在得到一组简化的解释值后，我们通过分析可以把这些解释属性分为三个基本类中的一种：



3、针对球赛统计分析系统

(1) 解释属性

在解释空间中的属性主要有：

- 场地：取值范围是好（通常指主场场地）、差（通常指客场场地）。
- 天气：取值范围是好、差。
- 裁判执法尺度：是指裁判在比赛中的判罚尺度，用于衡量裁判员是否按照规则精神来履行其职责和严格执法。裁判尺度对实力接近的胜负有着决定性的影响。取值范围是宽松、严格。
- 球员能力：分为我方球员的能力和对方球员的能力，其中我方球员的能力又可细分为防守能力、传球能力（是否到位）和控球能力。取值范围是强、弱。
- 我方球员的心态：取值范围是好、差。
- 我方前锋的状态：取值范围是好、差。
- 我方球员的拼抢程度：取值范围是大（拼抢凶狠）、小（拼抢不凶狠）。
- 我方球员的拦截次数：是指在比赛中总共拦截了多少次球。取值范围是多、少。
- 我方球员的关键传球次数：关键传球是指帮助队友完成射门的关键一传。关键传球的次数越多，可以为得分创造更多机会。这里主要是得到比赛中关键传球的次数。取值范围是多、少。

- 我方的技术型球员：取值范围是在场、不在场。

(2) 对应关系

要得到数据空间中一组偏差值对应到解释空间中的解释（多个偏差值解释的组合），首先要对每个偏差值进行分析，所用到的对应关系（规则）是：

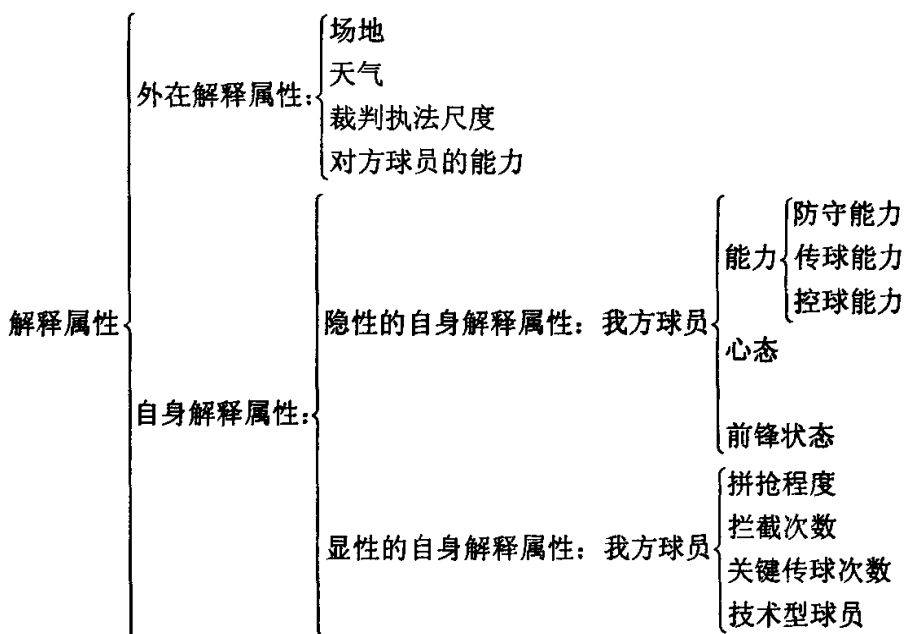
- 比赛分数大—裁判执法尺度宽松∨我方球员的能力强∨我方球员的关键传球次数多∨场地好
- 比赛分数小—裁判执法尺度严格∨我方球员的能力弱∨我方球员的关键传球次数少∨场地差
- 判罚次数大—裁判执法尺度严格∨我方球员的拼抢程度大∨我方球员的拦截次数多
- 判罚次数小—裁判执法尺度宽松∨我方球员的拼抢程度小∨我方球员的拦截次数少
- 控球时间大—场地好∨我方球员的拦截次数多∨我方球员的传球能力强∨我方技术型球员在场
- 控球时间小—场地差∨我方球员的拦截次数少∨我方球员的传球能力弱∨我方技术型球员不在场
- 我方射门次数大—天气好∨我方球员的控球能力强∨我方前锋的状态好∨我方球员的关键传球次数多
- 我方射门次数小—天气差∨我方球员的控球能力弱∨我方前锋的状态差∨我方球员的关键传球次数少
- 对方射门次数大—我方球员的防守能力弱∨对方能力强∨我方球员的心态差
- 对方射门次数小—我方球员的防守能力强∨对方能力弱∨我方球员的心态好
- 射正率大—我方前锋的状态好∨对方能力弱
- 射正率小—我方前锋的状态差∨对方能力强

然后将单独对每个偏差值分析得到的解释组合起来形成这组偏差值的一个整体解释，而所有的偏差值组合及其对应的解释形成了对应规则库。

通过球赛统计分析系统的对应规则库，我们可以找到这组偏差值的一个解释（由一组解释值构成）。然后，我们对这个解释进行多对多关系的处理（见第3.2.5部分），达到简化解释的目的。

(3) 解释属性的分类

在得到一组简化的解释值后，我们通过分析可以把这些解释属性分为三个基本类中的一种：



总之，在整个解释空间中所做的工作主要有：针对一个经过多对多关系处理的具体解释，在分类层次结构中进行搜索，从而得到这个解释所属的类别；然后，将按基本类划分的实际解释属性值与相应的正解释属性值表进行匹配，得到此基本类对这组实际偏差值的影响程度。这些工作可以用图3-3来表示。

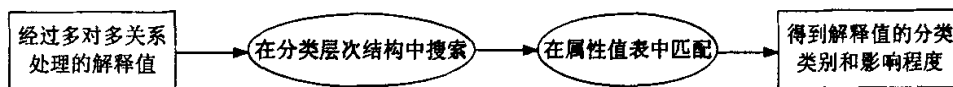


图3-3 解释空间中搜索与匹配

Figure 3-3 The search and match in Explanation Space

在接下来的3.2.3和3.2.4节中，我们将具体介绍这些工作和过程中所需要的标准，还有关于这些工作的更详细的图解。

3.2.3 解释标准与匹配

在数据空间和解释空间之间的联系是通过启发式建立起来的，可以用规则来表达，本文称之为对应关系。这种启发式关系往往由经验而来，建立过程中会忽略掉一些很难被察觉或很难被理解的中间关系。对于启发式分类问题，与一组观测值集合对应的是一个解决方案，而解决方案空间是一组预先定义好的类，即一组确定的解决方案；最终，一组观测值集合可以对应一个最优的解决方案，或是一组可接受的解决方案。而在面向解释的偏差分析中，与数据空间的一组偏差值集合对应的是一个由一组解释值构成的解释，而解释空间是通过由属性构建的分

类层次结构和由属性值构造的属性值表来描述的。由于我们进行偏差分析的目的是寻找产生偏差的原因，因此利用属性的分类，可以得到抽象层面的原因；再通过属性值表的匹配，区分原因所起的作用。

为了评价根据两个空间属性之间联系而得到的解释能力，本文给出一个解释标准：

设一组偏差值 $dev: ((f_1, v_1) (f_2, v_2) \cdots (f_n, v_n))$ ，一个解释 $exp: ((g_1, v_1') (g_2, v_2') \cdots (g_n, v_n'))$ ，则在匹配时将有：

- ① $(f_i, v_i) \in dev$ ，若与其对应的 $(g_j, v_j') \in exp$ ，则表示 g_j 是可解释的。
- ② $(f_i, v_i) \in dev$ ，若与其对应的 $(g_j, v_j') \notin exp$ ，而是 $(g_j, v_j'') \in exp$ ，则 g_j 是不匹配的 (v_j' 和 v_j'' 都是 g_j 的值)。
- ③ $(f_i, v_i) \in dev$ ，与其对应的是 (g_j, v_j') ，若 g_j 不是 exp 的一个属性，则表示 g_j 是不可解释的。
- ④ $(f_i, v_i) \notin dev$ ，与其对应的是 (g_j, v_j') ，若 g_j 是 exp 的一个属性，则表示 g_j 是缺失的。

应用在数据空间与解释空间之间建立联系的启发式规则（对应关系），可以根据一组偏差值得到由一组解释值构成的一个解释。在经过多对多关系处理后（见第3.2.5部分），针对这个解释，应用解释标准可以在解释空间中的分类层次结构和属性值表上进行匹配，但是要进行搜索得到解释的类别及影响程度，还需要有可进行比较的匹配标准。

3.2.4 匹配标准与搜索

在解释空间中，利用解释标准进行匹配时，为了得到抽象层次的原因解释及影响程度，还需要确定一个匹配标准，并用它引导在属性的分类层次结构上进行搜索及在属性值表上进行匹配。本文用 (I, E, U, M) 的形式来表示匹配标准，其中 I 表示不匹配的属性集合， E 是可解释的属性集合， U 是不可解释的属性集合， M 是缺失的属性集合。

3.2.4.1 分类层次结构上的搜索

将经过多对多关系处理的解释，在解释空间中的分类层次结构上进行搜索。在分类层次结构搜索后，得到这个解释所属的类别，这个类别将包括一些基本类，每个基本类都包括一些解释属性。在搜索过程中，主要以 U 和 M 为标准来进行比较。分类层次结构（见3.2.2节）中的每个节点分别代表一个类别，假定两个

节点 $S_1=(U_1, M_1)$ 和 $S_2=(U_2, M_2)$, S_1 好于 S_2 , 当且仅当:

$$(|U_1| < |U_2|) \vee ((|U_1| = |U_2|) \wedge (|M_1| < |M_2|))$$

其中 $|X_i| < |X_j|$ 的记法表明 X_i 集合所包含的元素比 X_j 集合少。

搜索是在解释空间中的分类层次结构自顶向下进行的:

- ①若顶层的节点好于下一层的所有子节点, 则顶层节点就是解;
- ②若顶层节点不比下一层子节点好, 则取最好的下一层子节点, 若最好的有多个, 则任取一个。若此时是底层节点, 则为解; 若不是底层节点, 则继续递归搜索。

搜索算法的形式化表示如下:

输入: 一个解释所含有解释属性

输出: 这个解释的类别

初始: S 表示一个节点。

将 S 初始化为分类层次结构中的顶层节点。

搜索: Search(S)

```
{
    if ( $S$  是底层节点)
         $S$  是最优解, 得到解释的类别, 返回;
    if ( $S$  好于所有直接子节点)
         $S$  是最优解, 得到解释的类别, 返回;
    Search( $S$  的最好直接子节点);
}
```

接下来, 针对三个具体实例来分析搜索的过程:

(1) 针对试卷成绩分析系统

①数据空间

一组偏差值:

峰度	偏度	实际平均分和预测平均分的关系
高狭峰	标准偏度	=

②对应关系

与一组偏差值对应的是一个解释 (由一组解释值构成)。经过多对多关系处理后, 解释变为:

试题梯度小 \wedge 学生能力分布接近 \wedge 学生素质分布接近

③解释空间

由这个解释,我们可以看出学生能力分布和学生素质分布是隐性的自身解释属性 b , 试题梯度是显性的自身解释属性 c 。然后,在解释空间中的分类层次结构中搜索:

步骤 1: 从顶层节点 S_1 ($\{abc\}$) 开始:

$$\because S_1 \text{ 中 } U_1 = \emptyset, M_1 = \{a\}$$

$$\text{又} \because S_1 \text{ 的第一个直接子节点 } S_2 (\{ab\}) \text{ 中 } U_2 = \{c\}, M_2 = \{a\}$$

$$\therefore |U_1| < |U_2|$$

$$\therefore S_1 \text{ 好于 } S_2$$

步骤 2: 将 S_1 的第二个直接子节点作为 S_2 ($\{ac\}$)

$$\because S_1 \text{ 中 } U_1 = \emptyset, M_1 = \{a\}$$

$$\text{又} \because S_2 \text{ 中 } U_2 = \{b\}, M_2 = \{a\}$$

$$\therefore |U_1| < |U_2|$$

$$\therefore S_1 \text{ 好于 } S_2$$

步骤 3: 将 S_1 的第三个直接子节点作为 S_2 ($\{bc\}$)

$$\because S_1 \text{ 中 } U_1 = \emptyset, M_1 = \{a\}$$

$$\because S_2 \text{ 中 } U_2 = \emptyset, M_2 = \emptyset$$

$$\therefore (|U_1| = |U_2|) \wedge (|M_1| > |M_2|)$$

$$\therefore S_1 \text{ 没有 } S_2 \text{ 好, 而且 } S_2 \text{ 是同层子节点中最好的节点}$$

步骤 4: 将 S_1 的第三个直接子节点 S_2 作为新的 S_1 ($\{bc\}$)

$$\because S_1 \text{ 中 } U_1 = \emptyset, M_1 = \emptyset$$

$$\text{又} \because S_1 \text{ 的第一个直接子节点 } S_2 (\{b\}) \text{ 中 } U_2 = \{c\}, M_2 = \emptyset$$

$$\therefore |U_1| < |U_2|$$

$$\therefore S_1 \text{ 好于 } S_2$$

步骤 5: 将 S_1 的第二个直接子节点作为 S_2 ($\{c\}$)

$$\because S_1 \text{ 中 } U_1 = \emptyset, M_1 = \emptyset$$

$$\because S_2 \text{ 中 } U_2 = \{b\}, M_2 = \emptyset$$

$$\therefore |U_1| < |U_2|$$

$$\therefore S_1 \text{ 好于 } S_2, \text{ 而且 } S_1 \text{ 好于其所有的直接子节点}$$

步骤 6: S_1 为最优解, 即 $\{bc\}$ 为最优解, 这个解释是第 7 类, 结束.

(2) 针对超市经营数据分析系统

①数据空间

一组偏差值:

客单价	客单数	平均毛利率	经营费用	利润
相等	相等	大	相等	大

②对应关系

与一组偏差值对应的是一个解释（由一组解释值构成）。经过多对多关系处理后，解释变为：

商品进价降低∧商品售价升高

③解释空间

从这个解释，我们可以看出商品进价和商品售价都是显性的自身解释属性 c 。然后，在解释空间中的分类层次结构中搜索：

步骤 1：从顶层节点 S_1 （{abc}）开始：

$\because S_1$ 中 $U_1 = \emptyset$, $M_1 = \{a, b\}$

又 $\because S_1$ 的第一个直接子节点 S_2 （{ab}）中 $U_2 = \{c\}$, $M_2 = \{a, b\}$

$\therefore |U_1| < |U_2|$

$\therefore S_1$ 好于 S_2

步骤 2：将 S_1 的第二个直接子节点作为 S_2 （{ac}）

$\because S_1$ 中 $U_1 = \emptyset$, $M_1 = \{a, b\}$

又 $\because S_2$ 中 $U_2 = \emptyset$, $M_2 = \{a\}$

$\therefore (|U_1| = |U_2|) \wedge (|M_1| > |M_2|)$

$\therefore S_1$ 没有 S_2 好

步骤 3：将 S_1 的第二个直接子节点 S_2 作为新的 S_1 （{ac}）

将未比较过的同层子节点（{bc}）作为 S_2

$\because S_1$ 中 $U_1 = \emptyset$, $M_1 = \{a\}$

又 $\because S_2$ 中 $U_2 = \emptyset$, $M_2 = \{b\}$

$\therefore (|U_1| = |U_2|) \wedge (|M_1| = |M_2|)$

$\therefore S_1$ 、 S_2 都是最好的，我们任选 S_1 作为同层子节点中最好的节点

步骤 4：将 S_1 的第一个直接子节点作为 S_2 （{a}）

$\because S_1$ 中 $U_1 = \emptyset$, $M_1 = \{a\}$

又 $\because S_2$ 中 $U_2 = \{c\}$, $M_2 = \{a\}$

$\therefore |U_1| < |U_2|$

$\therefore S_1$ 好于 S_2

步骤 5：将 S_1 的第二个直接子节点作为 S_2 （{c}）

$\because S_1$ 中 $U_1 = \emptyset$, $M_1 = \{a\}$

又 $\because S_2$ 中 $U_2 = \emptyset$, $M_2 = \emptyset$

$$\therefore (|U_1|=|U_2|) \wedge (|M_1|>|M_2|)$$

$\therefore S_1$ 没有 S_2 好, 而且 S_2 是同层子节点中最好的节点

步骤6: 将 S_1 的第二个直接子节点 S_2 作为新的 S_1

$\therefore S_1$ 是底层节点

$\therefore S_1$ 为最优解, 即 {c} 为最优解, 这个解释是第4类, 结束.

(3) 针对球赛统计数据分析系统

①数据空间

一组偏差值:

比赛 分数	判罚 次数	控球 时间	我方射 门次数	对方射 门次数	射正率
相等	相等	小	小	相等	相等

②对应关系

与一组偏差值对应的是一个解释 (由一组解释值构成)。经过多对多关系处理后, 解释变为:

场地差 \wedge 我方球员的拦截次数少 \wedge 我方球员的传球能力弱 \wedge 我方技术型球员不在场 \wedge 天气差 \wedge 我方球员的控球能力弱 \wedge 我方前锋的状态差 \wedge 我方球员的关键传球次数少

③解释空间

由这个解释, 我们可以看出天气、场地是外在的解释属性 a, 我方球员的传球能力、控球能力和我方前锋的状态是隐性的自身解释属性 b, 我方球员的拦截次数、我方的技术型球员和我方球员的关键传球次数是显性的自身解释属性 c。然后, 在解释空间中的分类层次结构中搜索:

步骤1: 从顶层节点 S_1 ({abc}) 开始:

$$\therefore S_1 \text{ 中 } U_1 = \emptyset, M_1 = \emptyset$$

$$\text{又} \therefore S_1 \text{ 的第一个直接子节点 } S_2 \text{ (}{ab}\text{)} \text{ 中 } U_2 = \{c\}, M_2 = \emptyset$$

$$\therefore |U_1| < |U_2|$$

$$\therefore S_1 \text{ 好于 } S_2$$

步骤2: 将 S_1 的第二个直接子节点作为 S_2 ({ac})

$$\therefore S_1 \text{ 中 } U_1 = \emptyset, M_1 = \emptyset$$

$$\text{又} \therefore S_2 \text{ 中 } U_2 = \{b\}, M_2 = \emptyset$$

$$\therefore |U_1| < |U_2|$$

$$\therefore S_1 \text{ 好于 } S_2$$

步骤3: 将 S_1 的第三个直接子节点作为 S_2 ({bc})

$\because S_1$ 中 $U_1 = \emptyset, M_1 = \emptyset$
 又 $\because S_2$ 中 $U_2 = \{a\}, M_2 = \emptyset$
 $\therefore |U_1| < |U_2|$
 $\therefore S_1$ 好于 S_2

步骤 4: \because 顶层节点 S_1 好于其所有的直接子节点

$\therefore S_1$ 为最优解, 即 {abc} 为最优解, 这个解释是第 8 类, 结束.

3.2.4.2 属性值表上的匹配

在分类层次结构搜索得到解释所属的类别后, 下一步就要在解释空间中的属性值表上进行匹配。不同于在分类层次结构搜索中以 U 和 M 为标准来比较, 在属性值表匹配时, 主要是以 I 和 E 为标准来进行比较, 其中 I 表示不匹配的属性集合, E 是可解释的属性集合。由一组偏差值得到一组解释值, 并将该组解释值按基本类进行区分, 然后用每个基本类中的解释属性及其正解释值来构成这个基本类的属性值表。最后将这个基本类中的解释属性的实际解释值与相应的属性值表进行匹配, 匹配结果将得到五种可能的情况, 分别代表这个基本类对所求问题的不同影响程度:

- ① 所有属性都是可解释的 ($I = \emptyset$), 称之为完全正解释。
- ② 所有属性都是不匹配的 ($E = \emptyset$), 称之为完全负解释。
- ③ 可解释的和 not 匹配的属性数相等 ($|E| = |I|$), 称之为平解释。
- ④ 可解释的属性数大于 not 匹配的属性数 ($|E| > |I|$), 称之为基本正解释。
- ⑤ 可解释的属性数小于 not 匹配的属性数 ($|E| < |I|$), 称之为基本负解释。

在属性值表上的匹配可以形式化地表示为:

输入: 一个解释所含有的一组实际解释值.

输出: 一组实际解释值所含有的每个基本类对所求问题的影响程度.

初始: B 表示一个基本类.

步骤 1: 将一组实际解释值所对应的解释属性按基本类进行区分, 得到 Num 个基本类, 每个基本类中含有 $Number$ 个解释属性.

步骤 2: 用每个基本类中的解释属性和其正解释值构成这个基本类的属性值表.

步骤 3: 在属性值表上进行匹配:

```

for (i=1; i<=Num; i++)
{
    Number=计算 $B_i$ 中的解释属性个数;
    for (j=1; j<=Number; j++)
    {

```

```

if( $B_i$ 中第 $j$ 个解释属性的实际解释值 $\in B_i$ 的属性值表)
    此解释属性是可解释的, 放到 $E_i$ 中;
else if( $B_i$ 中第 $j$ 个解释属性的正解释值 $\in B_i$ 的属性值表)
    此解释属性是不匹配的, 放到 $I_i$ 中;
}
if( $I_i = \emptyset$ )
     $M_i$ 是完全正解释, 得到此基本类的影响程度, 保存;
if( $E_i = \emptyset$ )
     $M_i$ 是完全负解释, 得到此基本类的影响程度, 保存;
if( $|E_i| = |I_i|$ )
     $M_i$ 是平解释, 得到此基本类的影响程度, 保存;
if( $|E_i| > |I_i|$ )
     $M_i$ 是基本正解释, 得到此基本类的影响程度, 保存;
if( $|E_i| < |I_i|$ )
     $M_i$ 是基本负解释, 得到此基本类的影响程度, 保存;
}

```

步骤4: 得到一组实际解释值所含有的每个基本类对所求问题的影响程度, 结束.

接下来, 针对三个具体实例进行分析:

(1) 针对试卷成绩分析系统

①数据空间

一组偏差值:

峰度	偏度	实际平均分和预测平均分的关系
高狭峰	标准偏度	=

②对应关系

与一组偏差值对应的是一个解释 (由一组解释值构成)。经过多对多关系处理后, 解释变为:

试题梯度小 \wedge 学生能力分布接近 \wedge 学生素质分布接近

③解释空间

由这个解释, 我们可以看出学生能力分布和学生素质分布是隐性的自身解释属性, 试题梯度是显性的自身解释属性。

经过前面的分类层次结构中的搜索这个解释已经划分为第7类, 也就是隐性的自身解释属性和显性的自身解释属性的组合。接下来, 我们将要分别构造这两个基本类的属性值表, 并详细说明二者在各自的属性值表

上的匹配过程及对这组偏差值的影响程度:

➤ 隐性的自身解释属性的影响程度

✧ 属性值表

因为含有学生的能力分布和学生的素质分布两个属性, 所以其正解释值所构成的属性值表为:

学生能力分布	接近
学生素质分布	接近

✧ 匹配

将实际的值与属性值表进行匹配:

∴ 偏差值:

(峰度, 高狭峰) \in 数据空间

又∴ 与其对应的解释值:

(学生能力分布, 接近)、(学生素质分布, 接近) \in 属性值表

∴ 学生能力分布是可解释的, 学生素质分布也是可解释的

∴ 隐性的自身解释属性是完全正解释

➤ 显性的自身解释属性的影响程度

✧ 属性值表

因为只含有试题梯度这一个属性, 所以其正解释值所构成的属性值表为:

试题梯度	大
------	---

✧ 匹配

将实际的值与属性值表进行匹配:

∴ 偏差值: (峰度, 高狭峰) \in 数据空间

又∴ 与其对应的解释值:

(试题梯度, 小) \notin 属性值表, 而是(试题梯度, 大) \in 属性值表

∴ 试题梯度是不匹配的

∴ 显性的自身解释属性是完全负解释

(2) 针对超市经营数据分析系统

① 数据空间

一组偏差值:

客单价	客单数	平均毛利率	经营费用	利润
相等	相等	大	相等	大

②对应关系

与一组偏差值对应的是一个解释（由一组解释值构成）。经过多对多关系处理后，解释变为：

商品进价降低∧商品售价升高

③解释空间

由这个解释，我们可以看出商品进价和商品售价都是显性的自身解释属性。经过前面的分类层次结构中的搜索这个解释已经划分为第4类，也就是显性的自身解释属性。接下来，我们将要构造这个基本类的属性值表，并详细说明它在属性值表上的匹配过程及对这组偏差值的影响程度：

➤ 显性的自身解释属性的影响程度

✧ 属性值表

因为含有商品进价和商品售价两个属性，所以其正解释值所构成的属性值表为：

商品进价	降低
商品售价	升高

✧ 匹配

将实际的值与属性值表进行匹配：

∴偏差值：（平均毛利率，大）∈数据空间

又∴与其对应的解释值：

（商品进价，降低）、（商品售价，升高）∈属性值表

∴商品进价是可解释的，商品售价也是可解释的

∴显性的自身解释属性是完全正解释

(3) 针对球赛统计数据分析系统

①数据空间

一组偏差值：

比赛分数	判罚次数	控球时间	我方射门次数	对方射门次数	射正率
相等	相等	小	小	相等	相等

②对应关系

与一组偏差值对应的是一个解释（由一组解释值构成）。经过多对多关系处理后，解释变为：

场地差∧我方球员的拦截次数少∧我方球员的传球能力弱∧我方技

术型球员不在场 \wedge 天气差 \wedge 我方球员的控球能力弱 \wedge 我方前锋的状态差 \wedge 我方球员的关键传球次数少

③解释空间

由这个解释，我们可以看出天气、场地是外在的解释属性，我方球员的传球能力、控球能力和我方前锋的状态是隐性的自身解释属性，我方球员的拦截次数、我方的技术型球员和我方球员的关键传球次数是显性的自身解释属性。

经过前面的分类层次结构中的搜索这个解释已经划分为第 8 类，也就是外在的解释属性、隐性的自身解释属性和显性的自身解释属性的组合。接下来，我们将要分别构造这三个基本类的属性值表，并详细说明三者各自的属性值表上的匹配过程及对这组偏差值的影响程度：

➤ 外在的解释属性的影响程度

✧ 属性值表

因为含有天气和场地两个属性，所以其正解释值所构成的属性值表为：

天气	好
场地	好

✧ 匹配

将实际的值与属性值表进行匹配：

∴ 偏差值：

(控球时间, 小)、(我方射门次数, 小) \in 数据空间

又∴ 与其对应的解释值：

(天气, 差)、(场地, 差) \notin 属性值表，而是(天气, 好)、(场地, 好) \in 属性值表

∴ 天气是不匹配的，场地是不匹配的

∴ 外在的解释属性是完全负解释

➤ 隐性的自身解释属性的影响程度

✧ 属性值表

因为含有我方球员的传球能力、控球能力和我方前锋的状态这三个属性，所以其正解释值所构成的属性值表为：

我方球员的传球能力	强
我方球员的控球能力	强
我方前锋的状态	好

◇ 匹配

将实际的值与属性值表进行匹配:

∴偏差值:

(控球时间, 小)、(我方射门次数, 小) ∈ 数据空间

又∴与其对应的解释值:

(我方球员的传球能力, 弱)、(我方球员的控球能力, 弱)、(我方前锋的状态, 差) ∉ 属性值表, 而是(我方球员的传球能力, 强)、(我方球员的控球能力, 强)、(我方前锋的状态, 好) ∈ 属性值表

∴我方球员的传球能力是不匹配的、控球能力是不匹配的、我方前锋的状态也是不匹配的

∴隐性的自身解释属性是完全负解释

► 显性的自身解释属性的影响程度

◇ 属性值表

因为含有我方球员的拦截次数、我方的技术型球员和我方球员的关键传球次数这三个属性, 所以其正解释值所构成的属性值表为:

我方球员的拦截次数	多
我方的技术型球员	在场
我方球员的关键传球次数	多

◇ 匹配

将实际的值与属性值表进行匹配:

∴偏差值:

(控球时间, 小)、(我方射门次数, 小) ∈ 数据空间

又∴与其对应的解释值:

(我方球员的拦截次数, 少)、(我方的技术型球员, 不在场)、

(我方球员的关键传球次数, 少) ∉ 属性值表, 而是(我方球员的拦截次数, 多)、(我方的技术型球员, 在场)、(我方球员的关键传球次数, 多) ∈ 属性值表

∴我方球员的拦截次数是不匹配的, 我方的技术型球员是不匹配的, 我方球员的关键传球次数也是不匹配的

∴显性的自身解释属性是完全负解释

在解释空间中,经历了分类层次结构上的搜索和属性值表上的匹配之后,最终结果将得到解释的类别(抽象层面的原因)和影响程度,以及具体的一组解释值。从以上可以看出:在解释空间中,利用解释标准和匹配标准在分类层次结构上进行搜索和在属性值表上进行匹配的具体过程可以描述成图3-4:

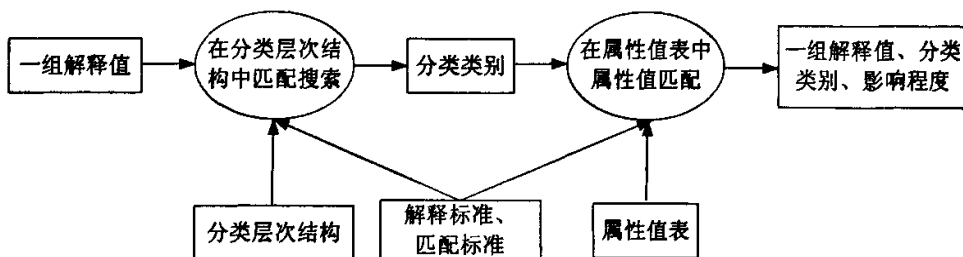


图 3-4 在解释空间中解释标准和匹配标准的使用

Figure 3-4 The using of explanation standard and matching standard in Explanation Space

3.2.5 多对多关系

数据空间中的一个属性的偏差值产生的原因可能关系到解释空间中的一个属性,也可能是多个属性;反过来,一个解释空间中的一个属性的取值可能是数据空间中的几个属性取值产生偏差的原因。通俗地讲,两个空间的属性取值之间的联系可能是多对多的关系。多个偏差值对应多个解释值的情况,可以分解为一对多和多对一的情况。实际上,多个偏差值对应一个解释值的情况所产生的问题可由一个偏差值对应多个解释值的情况来解决。对于一对多的情况,多个解释值应是一个启发式规则(对应关系)的结论部分,它是由解释值构成的析取式。对于这种析取式,若能修改成合取式,则可归为一对一的情况。因此,对于多对多的关系,可采用三种方法来处理:

- 1、若只有一对一的情况,则不对这种情况进行修改,直接使用解释标准和匹配标准在解释空间进行匹配、搜索。
- 2、若有一对多的情况,可分两种方法来处理:
 - ①对于多的方面,可通过获取实际解释属性的值进行确定,不一致的删除,一致的可以将析取改成合取。
 - ②对于多的方面,直接将析取改成合取,一般是将解释扩大化了,因此所得解释是近似的结果。

因为由一组偏差值根据对应规则库得到的一组初步解释值存在极大的不确定性,所以需要预先对这组初步解释值进行确定性处理,也就是多对多关系的处理。经过多对多关系的处理后,我们可以得到一组确定的具体解释值,从而便于在解释空间中的分类层次结构上和属性值表上进行搜索和匹配。

接下来，针对三个具体实例进行多对多关系处理的分析：

1、针对试卷成绩分析系统

(1) 多个偏差值对应多个解释值

高狭峰 \wedge 正偏度 \wedge 实际平均分和预测平均分相等

— (试题梯度小 \vee 学生能力分布接近 \vee 学生素质分布接近) \wedge (试题难度高 \vee 教学效果差 \vee 学生能力弱 \vee 学生素质低)

(2) 这种对应情况可以分解为一个偏差值对应多个解释值

高狭峰 — 试题梯度小 \vee 学生能力分布接近 \vee 学生素质分布接近
正偏度 — 试题难度高 \vee 教学效果差 \vee 学生能力弱 \vee 学生素质低

(3) 针对高狭峰和正偏度分别对应的多个解释属性

①如果实际可以获得是试题梯度小和试题难度高的原因，而不是其它原因，那么对应关系变为

高狭峰 \wedge 正偏度 \wedge 实际平均分和预测平均分相等

— 试题梯度小 \wedge 试题难度高

从而简化了解释，将复杂的析取式改成了合取式。

②如果无法确定是哪方面的原因，那么对应关系变为

高狭峰 \wedge 正偏度 \wedge 实际平均分和预测平均分相等

— 试题梯度小 \wedge 学生能力分布接近 \wedge 学生素质分布接近 \wedge 试题难度高 \wedge 教学效果差 \wedge 学生能力弱 \wedge 学生素质低

也就是说把所有的析取都变成合取，将解释扩大化了，结果是近似的。

2、针对超市经营数据分析系统

(1) 多个偏差值对应多个解释值

客单价大 \wedge 客单数大 \wedge 平均毛利率小 \wedge 经营费用相等 \wedge 利润相等

— (人民生活水平提高 \vee 商品售价降低) \wedge (人口增多 \vee 商品售价降低) \wedge (商品进价升高 \vee 商品售价降低)

(2) 这种对应情况可以分解为一个偏差值对应多个解释值

客单价大 — 人民生活水平提高 \vee 商品售价降低

客单数大 — 人口增多 \vee 商品售价降低

平均毛利率小 — 商品进价升高 \vee 商品售价降低

(3) 针对客单价、客单数和平均毛利率分别对应的多个解释属性

①如果实际可以获得均是商品售价降低的原因，而不是其它原因，那么对

应关系变为

客单价大 \wedge 客单数大 \wedge 平均毛利率小 \wedge 经营费用相等 \wedge 利润相等
—商品售价降低 \wedge 商品售价降低 \wedge 商品售价降低
—商品售价降低

从而简化了解释，将复杂的析取式改成了合取式。

②如果无法确定是哪方面的原因，那么对应关系变为

客单价大 \wedge 客单数大 \wedge 平均毛利率小 \wedge 经营费用相等 \wedge 利润相等
—人民生活水平提高 \wedge 商品售价降低 \wedge 人口增多
 \wedge 商品售价降低 \wedge 商品进价升高 \wedge 商品售价降低
—人民生活水平提高 \wedge 人口增多 \wedge 商品售价降低
 \wedge 商品进价升高

也就是说把所有的析取都变成合取，将解释扩大化了，结果是近似的。

3、针对球赛统计分析系统

(1) 多个偏差值对应多个解释值

比赛分数相等 \wedge 判罚次数相等 \wedge 控球时间小 \wedge 我方射门次数小 \wedge
对方射门次数相等 \wedge 射正率相等
—(场地差 \vee 我方球员的拦截次数少 \vee 我方球员的
传球能力弱 \vee 我方技术型球员不在场) \wedge (天气
差 \vee 我方球员的控球能力弱 \vee 我方前锋的状态差
 \vee 我方球员的关键传球次数少)

(2) 这种对应情况可以分解为一个偏差值对应多个解释值

控球时间小—场地差 \vee 我方球员的拦截次数少 \vee 我方球员的传球
能力弱 \vee 我方技术型球员不在场
我方射门次数小—天气差 \vee 我方球员的控球能力弱 \vee 我方前锋的
状态差 \vee 我方球员的关键传球次数少

(3) 针对控球时间和我方射门次数分别对应的多个解释属性

①如果实际可以获得是我方球员的传球能力弱和我方球员的关键传球次
数少的原因，而不是其它原因，那么对应关系变为

比赛分数相等 \wedge 判罚次数相等 \wedge 控球时间小 \wedge 我方射门次数小 \wedge
对方射门次数相等 \wedge 射正率相等
—我方球员的传球能力弱 \wedge 我方球员的关键传球
次数少

从而简化了解释，将复杂的析取式改成了合取式。

②如果无法确定是哪方面的原因，那么对应关系变为

比赛分数相等∧判罚次数相等∧控球时间小∧我方射门次数小∧
对方射门次数相等∧射正率相等

一场地差∧我方球员的拦截次数少∧我方球员的
传球能力弱∧我方技术型球员不在场∧天气差∧
我方球员的控球能力弱∧我方前锋的状态差∧我
方球员的关键传球次数少

也就是说把所有的析取都变成合取，将解释扩大化了，结果是近似的。

3.3 本章小结

为了使启发式分类方法能应用于更多的领域，同时为了能快速有效地帮助和改善人的解决问题能力和决策能力，我们提出了面向解释的偏差分析方法。它可纳入知识工程中启发式分类的范畴。

首先，介绍了面向解释的偏差分析的原理和模型化：采用不同于经典启发式分类的抽象方法给出了较为规范的抽象过程；采用属性和值分开的方法，通过应用解释标准在解释空间进行匹配；使用匹配标准在由解释属性构造的分类层次结构上搜索类别并在属性值表上区分影响程度，完成细化过程。最终得到带有抽象类别的人易理解的偏差解释。

其次，结合三个具体的领域来说明面向解释的偏差分析问题的模型化：试卷成绩分析领域、超市经营数据分析领域、球赛统计数据领域。通过针对这三个实例的详细分析，我们可以更深刻地理解这种模型化的过程。

第 4 章 面向解释的偏差分析问题解决方法

4.1 问题解决方法

在上一章中，我们通过以下概念来模型化面向解释的偏差分析问题：数据空间、解释空间、解释标准、匹配标准、分类层次结构、属性值表和多多对多关系。接下来，我们以此为基础总结出了面向解释的偏差分析问题解决方法。

面向解释的偏差分析问题解决方法描述了整个面向解释的偏差分析的过程：由原始数据在数据空间中抽象出偏差；将偏差对应到解释空间中的一个解释上；采用属性和值分开的方法，使用解释标准和匹配标准在解释空间中由解释属性构造的分类层次结构上搜索这个解释的类别并在属性值表上区分这个解释的影响程度，完成细化过程。其中，我们将与领域中的基础数据相对应的数值称为原始观测值和原始标准值。它们均可用集合来表示，集合的元素可表示成一组属性值对（由属性和值构成）的表。面向解释的偏差分析在输入了原始观测值和原始标准值后，通过上述的解决方法将得到关于产生偏差的解释类别和影响程度，以及具体的一组解释值。具体的问题解决方法（如图 4-1）可分为三个部分，共七个过程。

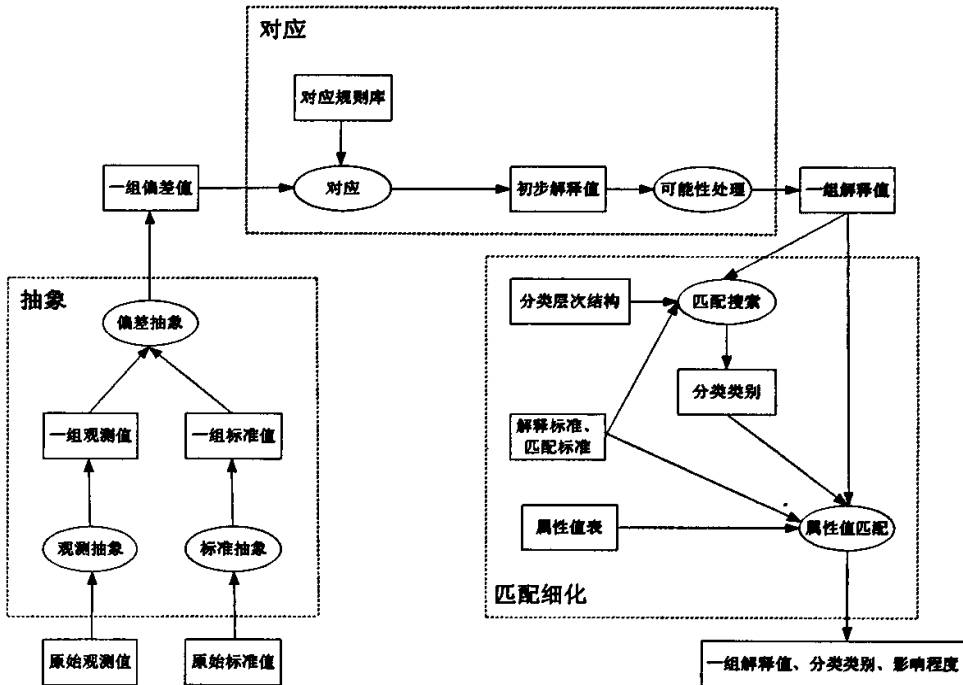


图 4-1 面向解释的偏差分析问题解决方法

Figure 4-1 The problem solving method of explanation-oriented deviation analysis

(1) 第一部分为抽象部分，包括观测抽象、标准抽象和偏差抽象过程。

①观测抽象：观测抽象是得到所需观测值的过程，主要包括选择抽象和应用抽象两个过程：选择抽象是在多个抽象方法中选择一个，应用抽象则是执行这个抽象方法。对于某些领域，最初输入的原始观测值是所需的观测值，则无需抽象。若不是所需的观测值，则需选择一个抽象方法，并执行它。

②标准抽象：标准抽象是得到所需标准值的过程，同样包括选择抽象和应用抽象两个过程。对于某些领域，最初输入的原始标准值是所需的标准值，则无需抽象。若不是所需的标准值，则需选择一个抽象方法，并执行它。

③偏差抽象：通过对观测值和标准值进行比较，得到偏差值，称之为偏差抽象。

在抽象部分，原始观测值、原始标准值的表示与获取，以及选择抽象的过程，最能体现出问题解决方法的 reusable 的特点。

(2) 第二部分为对应部分，包括对应和可能性处理过程。

①对应：输入数据空间的一组偏差值集合，根据对应规则库得到一个由一组解释值构成的解释。在数据空间和解释空间之间的联系是通过启发式建立起来的，用规则来表达，一个偏差值对应一个解释值或多个解释值，多个解释值表示成一条析取式作为规则的结论，多条规则组成对应规则库参与对应过程。

②可能性处理：也就是多对多关系的处理。若所有的规则都是一个偏差值对应一个解释值，则省略此步骤。若有规则是一个偏差值对应多个解释值，可采用两种方式来处理：一种是通过获取实际解释的值来与规则的结论进行匹配，若匹配失败，则删除此解释值，若匹配成功，则保留，并将析取式改成合取式。另一种是直接析取式改成合取式，一般是扩大了解释的范围。

(3) 第三部分为匹配细化部分，包括匹配搜索和属性值匹配过程。

①匹配搜索：输入的是一个解释的所有解释值的属性，利用解释标准在解释空间中以属性构建的分类层次结构上进行节点的匹配，并用可比较节点优劣的匹配标准进行搜索，找到可输出的解释的类别。

②属性值匹配：在得到了一个解释类别后，会按一些基本的类别再分类，然后将其中的每个基本的类别所包括的解释值与对应的解释属性及其特定值构成的属性值表进行匹配。匹配将可能得到五种情况，分别代表这个基本的类别对所求问题的不同影响程度。

4.2 基于本体实现面向解释的偏差分析

为了进一步实现知识的重用，我们以本体的形式来构造可重用的面向解释的

偏差分析系统,并通过 UPML 框架和 JESS 开发工具来实现这个系统。这是面向解释的偏差分析的第二部分工作,主要是由另一位同学负责,本人同时参与其中。

这部分工作的内容是在面向解释的偏差分析的前半部分工作基础上进行的,主要包括:首先,在面向解释的偏差分析模型化的基础上构建偏差分析本体;然后,应用问题解决方法统一描述语言 UPML^[49]构建面向解释的偏差分析问题解决方法框架,同时应用专家系统开发工具 JESS^[50]实现这一框架;最后,构建试卷成绩分析领域本体,并将这个可重用的面向解释偏差分析问题解决方法应用到试卷成绩分析的领域上,来说明一个通用的问题解决方法在针对某个具体领域时是如何实现的。

其实,在这部分工作中,我们主要用 UPML 的四个组件来构建面向解释的偏差分析问题解决方法组件库。UPML 是一个问题解决方法开发语言,它提供了人类可理解的问题解决方法的框架描述。UPML 定义了四个主要的知识组件:本体、任务、问题解决方法和领域模型。在我们的组件库中:

(1) 本体组件主要有:试卷成绩分析本体(STUDENT-SCORE-CONSTRUCT)、JESS 基础结构本体(JESS-BASE-CONSTRUCT)和偏差分析本体(DEVIATION-CONSTRUCT)。其中,偏差分析本体主要是定义了一组基本元素,包括原始观测值、原始标准值、观测值、标准值、偏差值、解释值、对应规则库、解释标准和匹配标准。任务、问题解决方法都要应用到它。

(2) 任务组件包括 9 个任务:偏差分析任务(deviation analysis-task)、抽象任务(abstractor-task)、对应任务(corresponding-task)、细化任务(refinemen-task)、观测抽象任务(ab-obs-task)、标准抽象任务(ab-nor-task)、偏差抽象任务(ab-dev-task)、层次匹配任务(ma-class-task)和值匹配任务(ma-value-task)。

(3) 问题解决方法组件包含 9 个主要的问题解决方法:偏差分析 PSM(deviation analysis-psm)、抽象 PSM(abstractor-psm)、对应 PSM(correspond-psm)、细化 PSM(match-psm)、观测抽象 PSM(ab-obs-psm)、标准抽象 PSM(ab-nor-psm)、偏差抽象 PSM(ab-dev-psm)、层次匹配 PSM(ma-class-psm)、值匹配 PSM(ma-value-psm)。

(4) 领域模型组件:引入了领域知识,包括被称为元知识的正确的事实与判断、知识库和被当作定理的中间结果,与具体实际领域有关。

为了将这些组件联系起来,我们还定义了多个桥来实现组件间的连接。通过这些桥,我们可以实现由领域数据到问题解决方法中所定义的和实例的映射。可以看出桥的定义可以使问题解决方法独立于具体的应用领域,从而实现问题解决方法的可重用性功能。

由于 UPML 仅仅是一种描述语言,其本身是不具备推理功能的,也就是说它是不能执行的。推理的实现依赖于所选择的对象语言。因此,我们在此选择对

象语言 JESS 来具体实现面向解释的偏差分析问题解决方法组件库，从而执行推理的过程。

4.3 本章小结

本章根据前一章中介绍的面向解释的偏差分析的模型化，总结出了面向解释的偏差分析问题解决方法。此方法描述了整个面向解释的偏差分析过程，主要分为三个部分，共七个过程：第一部分为抽象部分，包括观测抽象、标准抽象和偏差抽象过程；第二部分为对应部分，包括对应和可能性处理过程；第三部分为匹配细化部分，包括匹配搜索和属性值匹配过程。

此外，本章还简要介绍了另一位同学所完成的面向解释的偏差分析的第二部分工作，即以本体的形式来构造可重用的面向解释的偏差分析系统，并通过 UPML 框架和 JESS 开发工具来实现这个系统，最后将这个系统应用到试卷成绩分析的领域中。

结 论

面向解释的偏差分析以本体和问题解决方法为基础,通过计算观测值与标准值之间的偏差值并对其进行分析来寻找产生偏差的原因,从而给出了一个基于本体的偏差分析问题解决方法。它可以纳入知识工程中启发式分类的范畴。

本篇论文的内容就是对面向解释的偏差分析中前半部分工作的描述,重点通过一些基本概念来模型化面向解释的偏差分析问题,并通过试卷成绩、超市经营数据和球赛统计数据三个具体领域进行分析来详细说明了模型化的过程,最终总结出面向解释的偏差分析问题解决方法。为了模型化面向解释的偏差分析问题、并给出问题解决方法,在三年中我所做的工作具体包括:(1)分别对多个不同领域的实例数据进行面向解释的偏差分析的研究(2)设计并实现了其中一个面向解释的偏差分析系统,参与了其它面向解释的偏差分析系统的设计与实现工作。(3)通过对我们所做的其中三个有着相似特点的偏差分析系统进行比较、分析,最终抽象出一个比较通用的对偏差分析问题的形式化描述(模型化)和问题解决方法。在模型化面向解释的偏差分析问题时,我们采用属性和值分开的方法,将解释空间描述为由属性构造的分类层次结构和由属性值构造的属性值表,应用解释标准和匹配标准在分类层次结构上搜索解释类别、在属性值表上区分影响程度,最终得到带有抽象类别的人易理解的偏差解释。这是提出问题解决方法的前提,也是面向解释的偏差分析中后半部分工作的基础、构造本体的依据和前期准备工作。

面向解释的偏差分析与经典的启发式分类相比,既有相似之处,也有不同之处。相同的是,二者都有一个启发式的匹配过程,因此前者可以纳入后者的范畴、且适合在知识工程领域中加以探讨。不同之处是:

- ①与经典启发式分类需要较多的领域知识不同,面向解释的偏差分析主要是针对一些应用领域没有提供具有适宜分类层次的领域知识,并且对处理时间有较高要求的情况。因此,面向解释的偏差分析补充了启发式分类的类型,更具有实用性和时效性,处理问题更加方便快捷。
- ②所针对的通常是一些实际数据与预估结果有所偏差而需要解释的领域。因此,得到的解释有助于发现隐含的知识,从而资助和改善人们的解决问题能力和决策能力。
- ③此方法在后续工作中与本体的结合,使得我们的方法具有较高的可重用性和更普遍的应用价值,有利于知识的共享、重用和维护。

当然,面向解释的偏差分析也有一些缺陷,需要我们在以后的工作中不断的去完善它,它存在以下几个问题:

- ①其中有些发生可能性较小的或是不合理的偏差情况是被我们忽略的,有待进一步研究。
- ②我们只针对试卷成绩分析、超市经营数据分析、球赛统计数据领域进行了研究,虽然具有一定的通用性和实用性,但仍需要加强对更多实际领域的研究、扩展可重用领域,从而获得更易于人们理解的解释、使该方法更具有普遍性。

我们的最终目标是希望建立用于知识处理的语义 Web 服务,可以在知识的基础上对海量数据进行高水平的数据分析、数据校验和数据咨询等工作,从而实现一个利用已有知识来进行智能化信息处理的过程,有效地解决知识工程中知识的共享和重用问题。其中,面向解释的偏差分析作为数据分析的重要工作之一,目前已经在我和另一位同学的工作中得以实现。

在以后的工作中,我们将深入分析本体和问题解决方法的结合,克服以上提出的缺陷,不断改进和完善面向解释的偏差分析系统,从而使得系统可以更准确、更高效地解决实际应用中存在的大量问题。之后,我们还可以将面向解释的偏差分析系统在网作为我们语义 Web 服务中的一种服务进行发布,并在这个语义 Web 服务上添加其它的服务,从而具有更广泛的应用价值。

参考文献

- 1 Xiaohui Liu. Strategies for outlier analysis. Knowledge Discovery and Data Mining (Digest No. 1998/310), IEE Colloquium on 7 May 1998, Page(s): 3/1 - 3/3
- 2 Ahmet Kaya. Outlier Effects on Databases. In Proceedings of the third international conference on Advances in Information Systems(ADVIS 2004), Izmir, Turkey, 2004
- 3 Charu C. Aggarwal, Philip S. Yu. An effective and efficient algorithm for high-dimensional outlier detection. The VLDB Journal The International Journal on Very Large Data Bases, 2005, 14(2): 211-221
- 4 William J. Clancey. Heuristic Classification. Artificial Intelligence, 1985, 27(3): 289-350
- 5 Brandi Gilds. Knowledge engineering: ontology of world religions. Journal of Computing Sciences in Colleges, 2002, 18(1): 304-308
- 6 王鑫, 高小方. 面向实体的知识表示模型. 计算机科学, 2004, 第 31 卷(10 月增刊): 220-221, 271
- 7 王鑫, 韩成贵. 基于知识的数据校验. 计算机科学, 2004, 第 31 卷(10 月增刊): 282-284
- 8 V. Richard Benjamins and Dieter Fensel. Editorial: problem-solving methods. International Journal of Human Computer Studies, 1998, 49(4): 305-313
- 9 朱承, 曹泽文, 张维明. 知识库系统建模框架的发展与现状. 计算机工程, 2002, 28(8)
- 10 Sabine Bruaux, Gilles Kassel, Gilles Morel. An ontological approach to the construction of problem-solving models. K-CAP 2005: 181-182
- 11 Sabine Bruaux and Gilles Kassel. OntoKADS: a core ontology for developing expertise models for the CommonKADS methodology. In Proceedings of the EKAW*04 Workshop on Core Ontologies in Ontology Engineering, Northamptonshire (UK), 2004
- 12 B. Chandrasekaran, J. R. Josephson and V. Richard Benjamins. Ontology of Tasks and Methods. In: Proceedings of the Workshop on Knowledge Acquisition, Modeling and Management (KAW'98), Banff, Canada, 1998
- 13 V. Richard Benjamins and Asunción Gómez-Pérez. Knowledge-system Technology: Ontologies and Problem-Solving Methods[EB/ OL]. <http://hcs.science.uva.nl/usr/richard/pdf/kais.pdf>, 2004
- 14 Asunción Gómez-Pérez and V. Richard Benjamins. Overview of Knowledge

- Sharing and Reuse Components: Ontologies and Problem-Solving Methods. In Proceedings of the IJCAI-99 workshop on Ontologies and Problem-Solving Methods (KRR5), Stockholm, Sweden, 1999: 1-15
- 15 钱巧能, 刘亚军. Ontology 在 Semantic Web Services 中的应用综述. 计算机与数字工程, 2005, 33(11)
- 16 陆汝钤. 世纪之交的知识工程与知识科学. 北京: 清华大学出版社, 2001
- 17 谢储晖, 郭达志. 本体的研究现状及其应用. 福建电脑, 2003, (12)
- 18 Ying Ding. Ontology Research and Development: Part 2-A Review of Ontology Mapping and Evolving. Journal of Information Science, 2002, 28(5): 375-388
- 19 Ying Ding. Ontology Research and Development: Part 1-A Review of Ontology Generation. Journal of Information Science, 2002, 28(2): 123-136
- 20 Noy F.N., McGuinness D.L.. Ontology Development 101: A Guide to Creating Your First Ontology. Stanford Knowledge Systems Laboratory Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001-0880, 2001
- 21 陈壮生, 瞿裕忠. 基于本体的信息处理系统的设计与实现. 计算机工程, 2005, 31(11)
- 22 张泰. 基于本体的知识表示和知识共享. 江苏经贸职业技术学院学报, 2006, (1)
- 23 赵丽. 本体的理论及其应用研究. 兰州理工大学硕士学位论文, 2006. 4
- 24 林夕伟. 基于本体论的 E-MAIL 分类方法. 浙江教育学院学报, 2005, (2)
- 25 史忠植. 知识发现. 北京: 清华大学出版社, 2002
- 26 Xiaohui Liu, Gongxian Cheng and John X. Wu. Analyzing Outliers Cautiously. Knowledge and Data Engineering, IEEE Transactions on March-April 2002, 14(2): 432-437
- 27 Jacques Goupy. Outliers and experimental designs. Chemometrics and Intelligent Laboratory Systems, December, 1996, 35(2): 145-156
- 28 Wen Jin, Yuelong Jiang, Weining Qian, and Anthony K.H. Tung. Mining Outliers in Spatial Networks. In Proceedings of the 11th International Conference on Database Systems for Advanced Applications (DASFAA'2006), Singapore, 2006
- 29 Jiaqi Wang, Chengqi Zhang, Xindong Wu, et al.. SVM-OD: SVM Method to Detect Outliers. Foundations and Novel Approaches in Data Mining. Springer Berlin / Heidelberg, 2005: 129-141
- 30 安淑芝等. 数据仓库与数据挖掘. 清华大学出版社, 2005. 6
- 31 张云涛, 龚玲. 数据挖掘原理与技术. 电子工业出版社, 2004. 4

- 32 武森, 高学东. 数据仓库与数据挖掘. 冶金工业出版社, 2003. 9
- 33 Dantong Yu, Gholamhosein Sheikholeslami and Aidong Zhang. FindOut: Finding Outliers in Very Large Datasets. Knowledge and Information Systems, 2002, 4(4): 387-412
- 34 Ana-Maria-Marhan, Fabio Paternò, Carmen Santoro. Improving Dependability through a Deviation Analysis on Distributed Tasks in Safety Critical Systems. The International Conference on Dependable Systems and Networks, Palazzo dei Congressi, Firenze, Italy, 2004
- 35 Knorr E, Ng R. Algorithms for mining distance-based outliers in large datasets. In Proceedings of the 24th VLDB conference, New York, 1998, pp: 392-403
- 36 Knorr E, Ng R. Finding intensional knowledge of distance-based outliers. In Proceedings of 25th VLDB conference, Edinburgh, 1999, pp: 211-222
- 37 王新. 浅析人工智能与知识工程的关系. 云南民族学院学报(自然科学版), 2000, 9(2): 115-117
- 38 黄日茂. 语义 Web 知识表示方法的研究. 贵州大学硕士学位论文, 2006. 6
- 39 Dieter Fensel and Enrico Motta. Structured development of problem solving methods. IEEE Transactions on Knowledge and Data Engineering, page to appear, 2000
- 40 William J. Clancey. Model Construction Operators. Artificial Intelligence, 1992, 53(1): 1-115
- 41 Mark J Stefik. Introduction to Knowledge Systems. San Francisco, CA: Morgan Kaufmann, 1995
- 42 B. J. Wielinga, J. M. Akkermans and A. Th. Schreiber. A Competence Theory Approach to Problem Solving Method Construction. International Journal of Human-Computer Studies, 1998, 49(4): 315-338
- 43 K.-D. Althoff, M. Manago, R. Bergmann, F. Maurer, S. Wess, E. Auriol, N. Conruyt R. Traphöner, M. Bräuer. Induction and Case Based Reasoning for Classification Tasks. In: H. H. Bock, W. Lenski and M. M. Richter (eds.), Information Systems and Data Analysis, Prospects-Foundations-Applications, Proceedings of the 17th Annual Conference of the GfKI, University of Kaiserslautern, Springer Verlag, Berlin-Heidelberg, 1994: 3-16
- 44 Enrico Motta and Wenjin Lu. A library of components for classification problem solving. In: Pacific Rim Knowledge Acquisition Workshop, Sydney, Australia, 2000
- 45 Enrico Motta, John Domingue, Liliana Cabral and Mauro Gaspari. IRS-II: A

- Framework and Infrastructure for Semantic Web Services. In Proceedings of the 2nd International Semantic Web Conference (ISWC2003), Sundial Resort, Sanibel Island, Florida, USA, 2003
- 46 贾俊平, 金勇进. 统计学. 北京: 中国人民大学出版社, 2004
- 47 胡学锋. 统计学. 广州: 中山大学出版社, 1999
- 48 侯国平. 提高门店赢利能力的控制点. <http://www.mie168.com/manage/2005-01/15288.htm>, 2005-1-25
- 49 Dieter Fensel, Enrico Motta, V. Richard Benjamins, et al.. The Unified Problem-solving Method Development Language UPML. Knowledge and Information Systems, 2003, 5(1): 83-131
- 50 Ernest Friedman-Hill. Jess in Action: Java Rule-Based Systems. Greenwich: Manning, 2003

攻读硕士学位期间发表的学术论文

- 1 马玉莲, 王宇冬, 王鑫. 基于解释的分类算法. 全国第 18 届计算机技术与应用 (CACIS • 2007) 学术会议, 浙江宁波, 2007. 安徽, 中国科技大学出版社 (已录用)
- 2 王鑫, 王宇冬, 马玉莲. 面向解释的偏差分析问题解决方法. 计算机研究与发展 (增刊) (已录用)

致 谢

本论文的研究工作是在王鑫老师的悉心指导下完成的。从论文选题到最终定稿，王老师都给予了极大的关怀、支持和指导。

首先由衷地感谢王鑫老师三年来的培养和教育，以及在生活上的关心和帮助。作为王鑫老师的研究生，他对我在研究上的指导，在学业上的督促和辅导，在思想上的教育，使我受益匪浅。

同时，感谢刘椿年、刘建丽、戴伟长、冀俊忠、邓米克、蒋强荣等老师，他们严谨的研究作风、朴实的生活态度，都值得我去学习。

另外，向王宇冬、禹晶、杨雅媛、孙守霞、王小艳、韩婷等同学这三年来在学业和生活上对我的帮助表示感谢。在此，还要感谢参加了本项研究工作的北京工业大学的一些本科学生。

最后感谢各位参加评审和参加答辩会的老师，感谢您们在百忙之中抽出时间来对我进行指导和帮助。