《数据挖掘》

课程实验教学指导书

课程编号：

撰写人：李春林

审核人：

河北经贸大学

数学与统计学学院

2010 年 6 月 30 日

前　　言

一、实验总体目标

《数据挖掘》实验大纲，立足于全面，系统地反映数据挖掘分析问题和解决问题的基本思想和经典方法，结合本学科最新理论研究成果来组织编写，使学生能系统正确地掌握数据挖掘的理论基础，掌握几种基本的数据挖掘方法，培养学生初步具有能结合实际情况对具体项目进行数据收集和对所获得数据进行处理和进行数据挖掘操作的能力。

二、适用专业年级

统计学、数学与应用数学、信息与计算科学专业的三年级学生

三、先修课程

概率论与数理统计、多元统计分析、时间序列分析

四、实验项目及课时分配

实验项目		实验要求	实验类型	每组人数	实验学时
实验一	基本操作	必修	验证性	1	2
实验二	数据选择	必修	验证性	1	2
实验三	数据预处理	必修	验证性	1	3
实验四	关联规则挖掘	必修	验证性	1	2
实验五	分类：决策树	必修	验证性	1	2
实验六	因子分析	必修	验证性	1	2
实验七	聚类分析	必修	验证性	1	2
实验八	神经网络	必修	验证性	1	2

五、实验环境

计算机、Spss-Clementine 数据挖掘软件

六、实验总体要求

1．熟练应用 Spss-Clementine 数据挖掘软件，使之尽量具有通用性。

2．上机前充分准备，复习有关方法，实际操作并反复查对操作过程，列出上机步骤。

3．学会使用 SPSS 软件进行非参数统计分析。

4．完成计算后写出计算实验报告，内容包括：各种方法的实际应用，数据预处理方法

说明，变量说明，输出分析结果，结果分析、小结、备注等。

七、本课程实验的重点、难点及教学方法建议

重点：熟练掌握数据挖掘技术的各种分析方法

难点：能够将所学的各种方法进行实例分析，对数据集当中的缺失值进行合理的数据预处理，以及对结果合理的解释。

教学方法建议：教师采用课堂讲授，配合上机练习巩固所学内容的教学方法，要求学生重点掌握当今世界上流行数据挖掘技术的各种分析方法，对数据挖掘有一个总体的认识，重点掌握 Spss-Clementine 数据挖掘软件的实际操作。

实验一：基本操作

一、实验目的

1、理解对大型的、复杂的和信息丰富的数据集进行分析的必要性；

2、明确数据挖掘过程的目标和首要任务；

3、描述数据挖掘技术的起源；

4、了解数据挖掘软件 Spss-clementine 的基本功能。

二、实验内容

1、数据挖掘概述；

2、数据挖掘的起源；

3、数据挖掘的过程；

4、数据挖掘软件 Spss-Clementine 的基本功能和操作。

三、实验仪器设备和材料清单

1、计算机；

2、Spss-Clementine 数据挖掘软件。

四、实验要求

教师采用课堂讲授，配合上机练习巩固所学内容的教学方法，要求学生重点掌握了解当今世界上流行数据挖掘技术的应用范围和流程，对数据挖掘有一个总体的认识，重点掌握 Spss-Clementine 的基本功能和操作。

五、实验过程

（一）数据挖掘概述：

1．数据化信息产业的发展引发了数据的大量聚集，而如何将这些数据转化成有用有信息和知识是信息领域所面临的问题。

2．在数据库开发设计中经历了二个阶段的演化：

第一阶段：数据收集和数据库创建，数据管理引发了数据存储和检索，数据库事务处理技术研究。

第二阶段：数据分析与理解引发了数据仓库和数据挖掘技术的研究。

数据仓库和数据挖掘技术的出现从根本上是为了解决这样一个问题：在创建一个数据集，考虑数据的存储效率的时候，同时考虑数据最终如何被使用和分析。

例如，数据收集和数据库创建机制为数据存储和检索、查询和事务处理有效机制开发的必备基础。随着提供查询和事务处理的大量数据库系统（如医院中使用的各种信息系统）广泛应用，数据分析和理解自然成为下一个目标。

3．数据挖掘的两个根本目标：预测和描述

预测涉及到使用数据集中的一些变量或域来预测其他我们关心的变量的未知或未来的值；描述关注的则是找出描述可由人类解释的数据格式。

1)预测性数据挖掘:生成已知数据集的系统模型。

2)描述性数据挖掘:在数据集上生成新的、非同寻常的信息。

4．数据挖掘的基本任务：

1)分类：

2)回归：

3)聚类：

4)总结概括：

5)关联建模：

6)变化与偏差检测：

（二）数据挖掘的起源

1．大部分数据挖掘问题和相应的解决方法都起源于传统的数据分析。

2．数据挖掘起源于多种学科，主要是统计学和机器学习。

3．统计学起源于数学，它强调数据上精确；机器学习主要起源于计算机实践，它侧重于对事物的检验，确定它表现的好坏。

4．数据挖掘中的基本模型法则起源于控制理论，控制理论主要应用于工程系统和工业过程。

5．在控制理论中通过观察一个未知系统的输入输出信息，来决定其数学模型的问题常被称为系统识别。

6．系统识别是多样化的，从数据挖掘的立场出发是预测系统的行为，并解释系统变量之间的相互作用和关系。

（三）数据挖掘的过程

1．定义：数据挖掘是一个从已知数据集合中发现各种模型、概要和导出值的过程。

2．陈述问题和阐明假设

大多数基于数据的模型研究都是在一个特定的应用领域里完成的。为了提出一个有意义的问题的陈述，需要拥有该领域内丰富的知识和经验，着重对问题的清晰描述，而不是过分关注数据挖掘技术。尽可能地为未知的相关性指定一组变量，指定此相关性的一个大体形式作为初始假设。这一步需要应用领域的专门技术与数据挖掘模型相结合。

3．数据收集

数据如何收集，有两种截然不同的可能：

1）当数据产生过程在专家的控制下时，称为“设计实验”。

2）专家不能影响数据产生过程，称为“观察法”，数据随机产生。

通常收集完成后取样的分布也是完全未知的，或者是在数据收集过程中部分或者不明确地给出，但要理解数据收集是怎样影响它的理论分布的，这一点相当重要。

4．数据预处理

数据常常采集于已有的数据库、数据仓库和数据集市中。数据预处理有两个任务：

1）异常点的检测(和去除)：异常点是与众不同的数值，它们与大多数观察值不一致。

2）比例缩放、编码和选择特征：数据预处理包括各种比例缩放和不同类型的编码。

例如：取[1，0]的特征和取[-100，100]的特征，其加权值是不一样的，对数据挖掘的结果的影响也不尽相同。因此进行比例缩放使它们的加权相同。

5．模型评估

选择并实现适当的数据挖掘技术是这一步骤的主要任务。在应用中，建立在几个模型的基础上的，从中选择最好的模型是额外的任务。了解从数据中学习和发掘的基本原则，并掌握一些特殊的技术，应用这些技术可以从数据中成功在学习，也可以应用这些的技术找到适当的模型，这些内容将在相关章节的获得。

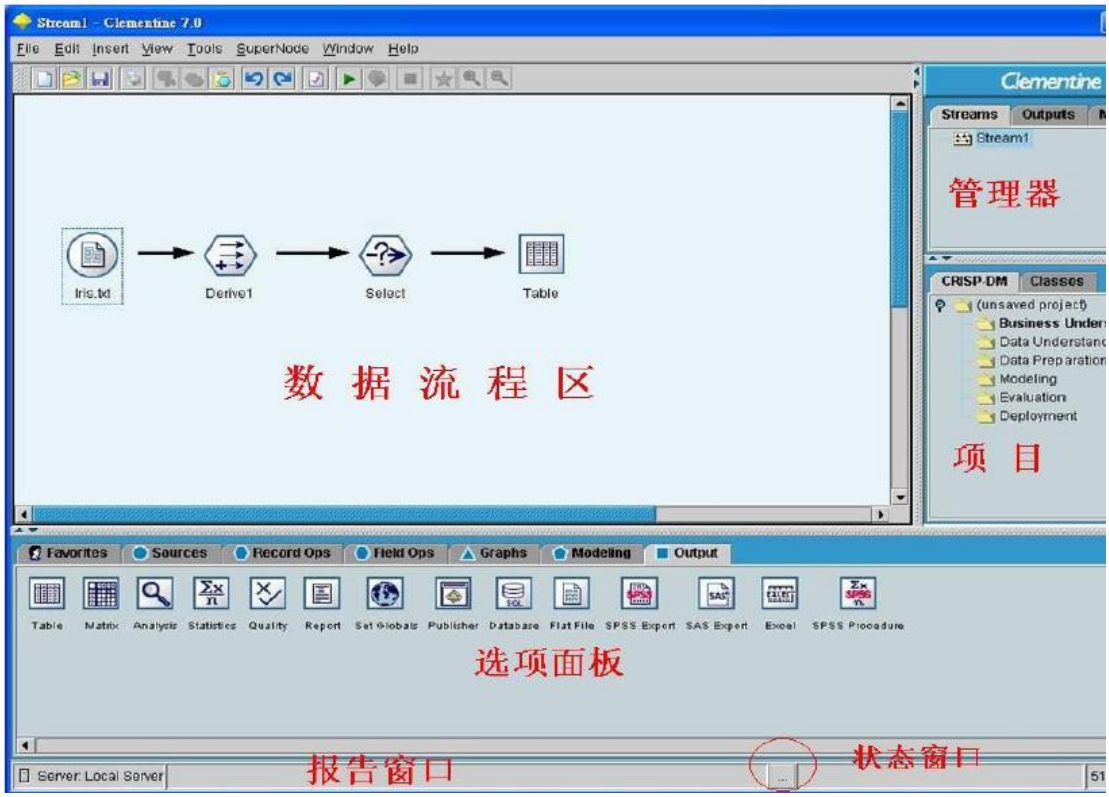
6．解释模型和得出结论

在大多数应用场合，数据挖掘模型应该有助于决策。一般来说，简单的模型容易说明，但其准确性就差一些，现代的数据挖掘方法着重于使用高维度的模型来获得高精度的结果。

用特定的技术验证这些结果对这些模型进行解释说明被认为是一项独立的任务，同时也是非常重要的。

（四）数据挖掘软件 Spss-Clementine 的基本功能

1．操作界面的介绍



1)数据流程图

Clementine 在进行数据挖掘时是基于数据流程形式，从读入数据到最后的结果显示都是由流程图的形式显示在数据流程区内。数据的流向通过箭头表示，每一个结点都定义了对数据的不同操作，将各种操作组合在一起便形成了一条通向目标的路径。

数据流程区是整个操作界面中最大的部分，整个建模过程以及对模型的操作都将在这个区域内执行。我们可以通过 File－new stream 新建一个空白的数据流，也可以打开已有的数据流。所有在一个运行期内打开的数据流都将保存在管理器的 Stream 栏下。

2)选项面板

选项面板横跨于 Clementine 操作界面的下部，它被分为 Favorites、Sources、Record Ops、Fields Ops、Graphs、Modeling、Output 七个栏，其中每个栏目包含了具有相关功能的结点。

结点是数据流的基本组成部分，每一个结点拥有不同的数据处理功能。设这不同的栏式为了将不同功能的结点分组，线面我们介绍各个栏的作用。

Sources：该栏包含了能读入数据到 Clementine 的结点。例如 Var.File 结点读取自由格式的文本文件到 Clementine，Spss File 读取 Spss 文件到 Clementine。

Record Ops：该栏包含的结点能对数据记录进行操作。例如筛选出满足条件的记录（select）、将来自不同数据源的数据合并在一起（merge）、向数据文件中添加记录（append）等。

Fields Ops：该栏包含了能对字段进行操作的结点。例如过滤字段（filter）能让被过滤的字段不作为模型的输入、derive 结点能根据农户定义生成新的字段，同时我们还可以定义字段的数据格式。

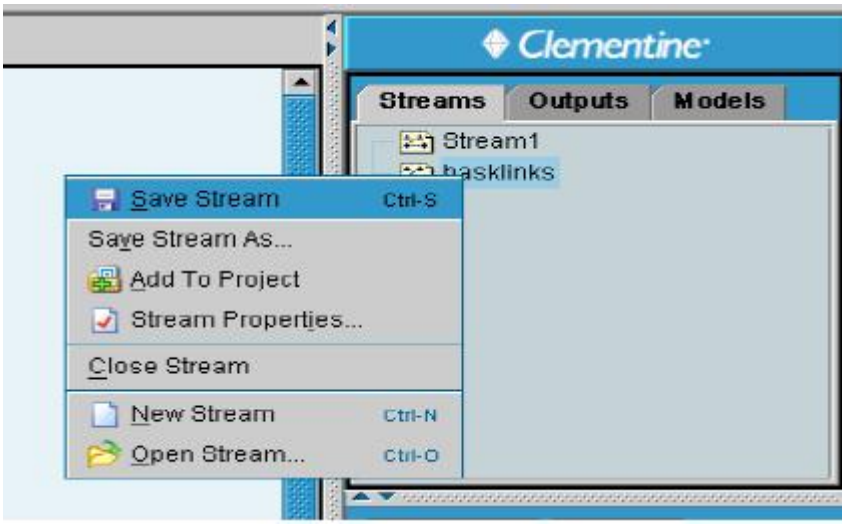
Graphs：该栏包含了纵多的图形结点，这些结点用于在建模前或建模后将数据由图形形式输出。

Modeling：该栏包含了各种已封装好的模型，例如神经网络（Neural net）、决策树（C5.0）等。这些模型能完成预测（Neural net，Regression，Logistic）、分类（C5.0，C&R Tree，Kohonen，K-means，Two-step）、关联分析（Apriori，GRI，Sequece）等功能。

Output：该栏提供了许多能输出数据、模型结果的特点，农户不仅可以直接在 Clementine 中查看输出结果，也可以输出到其他应用程序中查看，例如 Spss 和 Excel。

Favorites：该栏放置了用户经常使用的结点，方便用户操作。用户可以自定义其 Favorites 栏，操作方法为：选中菜单栏的 Tools，在下拉菜单中选择 Favorites，在弹出的 Palette Manager 中选中要放入 Favorites 栏中的结点。

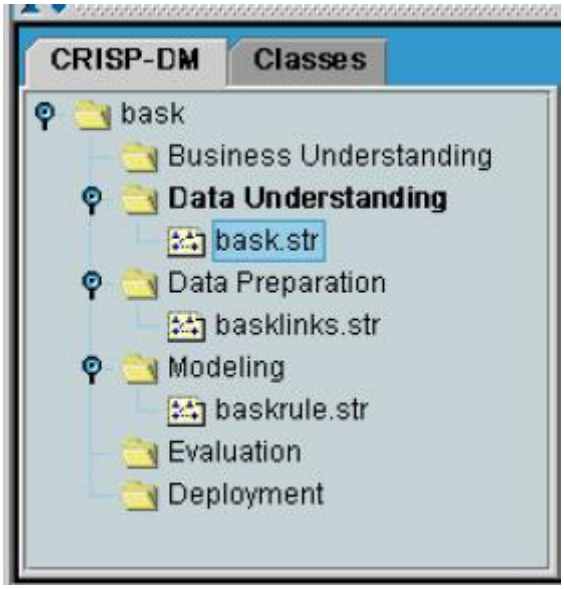
3)管理器



管理器中共包含了 Streams、Outputs、Models 三个栏。其中 Streams 中放置了运行期内打开的所有数据流，可以通过右键单击数据流名对数据流进行保存、设置属性等操作。Outputs 中包含了运行数据流时所有的输出结果，可以通过双击结果名查看输出的结果。Models 中包含了模型的运行结果，我们可以右键单击该模型从弹出的 Browse 中查看模型结果，也可以将模型结果加入到数据流中。

4)项目窗口的介绍

项目窗口含有两个选项栏，一个是 CRISP-DM，一个是 Classes。



CRISP-DM 的设置是基于 CRISP-DM Model 的思想，它方便用户存放在挖掘各个阶段形成的文件。由右键单击阶段名，可以选择生成该阶段要拥有的文件，也可以打开已存在的文件将其放入该阶段。这样做的好处是使用户对数据挖掘过程一目了然，也有利于对它进行修改。

Classes 窗口具有同 CRISP-DM 窗口相似的作用，它的分类不是基于挖掘的各个过程，而是基于存储的文件类型。例如数据流文件、结点文件、图表文件等。

六、实验报告要求

1、能够将分析过程中的软件操作步骤在报告中凸现出来；

2、结果的分析要合理、准确。

七、思考题

1、为什么数据挖掘者对数据的理解很重要；

2、口述 Spss-Clementine 数据挖掘界面每个窗口的基本功能；

3、了解 SAS 数据挖掘软件的基本操作流程。

八、注意事项

1、数据挖掘技术的全面认识；

2、软件操作应当正确且熟练掌握。

实验二　数据选择

一、实验目的

1、分析原始大型数据集的基本表述和特征。

2、对数值型属性应用不同的标准化技术。

3、掌握技术数据挖掘软件 Spss-Clementine 的数据流的创建与操作。

二、实验内容

1、原始数据的表述；

2、原始数据的特性；

3、数据流基本操作的介绍；

三、实验仪器设备和材料清单

1、计算机；

2、Spss-Clementine 数据挖掘软件。

四、实验要求

教师采用课堂讲授，配合上机练习巩固所学内容的教学方法，要求学生重点掌握数据的选择，能正确解释软件处理的结果，尤其是样本信息的解释；同时要求学生阅读一定数量的文献资料，掌握软件的实际操作。

五、实验过程

（一）原始数据的表述

1．常见的数据类型

数据挖掘过程的基本对象是数据样本，每个样本都用几个特征来描述，每个特征有不同的类型的值。常见类型：数值型和分类型。数值型的值包括实型变量和整型变量。

1)数值型：其特征是其值有顺序关系和距离关系。

2)分类型：其特征是变量间是否相等，且可用二进制数来表述。

2．基于变量值的变量分类法：连续型变量和离散型变量

1)连续型变量也称为定量型或度量型变量。可用间隔尺度或比例尺度来衡量。温度尺度属间隔尺度，没有绝对零点。高度、长度和工资属比例尺度，有绝对零点，

2)离散型变量也称为定性型变量。可用名义尺度或有序尺度来衡量。顾客类型标志和邮编属名义尺度，排名属有序尺度。

3．基于数据的与时间有关的行为特性的类型：静态数据和动态数据

（二）原始数据的特性

在数据挖掘初始阶段面对的数据也许有潜在的杂乱性，存在着丢失值、失真、误记录和不适当的样本。因此在必须根据已有的数据甚至是缺失值的数据进行建模。这样就可能避免在挖掘前处理缺失值问题。

另一个问题是必须有处理“非常值”的机制，来消除“非常值”对最终结果的影响，数据可能并不是来自我们假定的总体。异常点是典型的例子。失真的数据、方法上错误的步骤、滥用挖掘工具、模型太理想化、超出各种不确定性和模糊性的数据来源的模型可能导致挖掘方向的错误。因此挖掘不只是简单在应用一系列工具于已知问题，而是一种批判性的鉴定、考查、检查以及评估过程。

1．挖掘过程中一个最关键的步骤是对初始数据集的预备和转换，数据预备有两个中心任务：

1)把数据组织成一种标准形式，使其能被挖掘工具和其他基于计算机的工具处理（一个关系表）

2)准备数据集使之能得到最佳的挖掘效果

（三）数据流基本操作的介绍

1) 生成数据流的基本过程

数据流是由一系列的结点组成，当数据通过每个结点时，结点对它进行定义好的操作。我们在建立数据流是通常遵循以下四步：

① 向数据流程区增添新的结点；

② 将这些结点连接到数据流中；

③ 设定数据结点或数据流的功能；

④ 运行数据流。

2) 向数据流程区添/删结点

当向数据流程区添加新的结点时，我们有下面三种方法遵循：

① 双击结点面板中待添加的结点；

② 左键按住待添加结点，将其拖到数据流程区内；

③ 选中结点面板中待添加的结点，将鼠标放入数据流程区，在鼠标变为十字形时单击数据流程区。

通过上面三种方法我们都将发现选中的结点出现在了数据流程区内。

当我们不再需要数据流程区内的某个结点时，可以通过以下两种方法来删除：

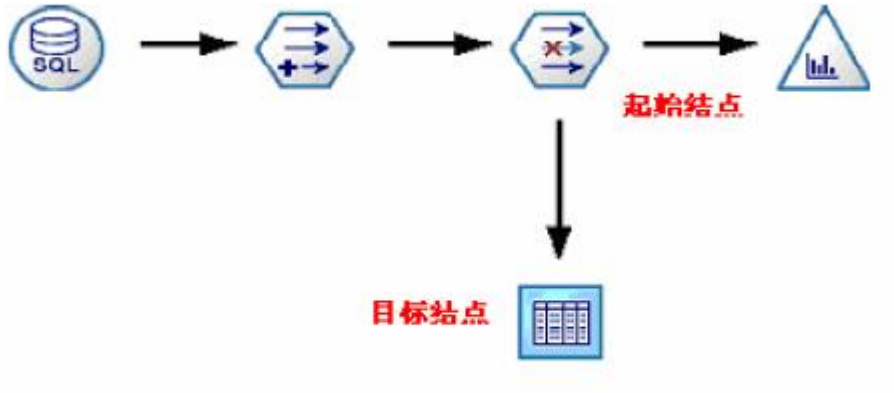
① 左键单击待删除的结点，用 delete 删除；

② 右键单击待删除的结点，在出现的菜单中选择 delete。

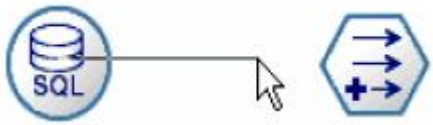
3）将结点连接到数据流中

上面我们介绍了将结点添加到数据流程区的方法，然而要使结点真正发挥作用，我们需要把结点连接到数据流中。以下有三种可将结点连接到数据流中的方法：

① 双击结点。左键选中数据流中要连接新结点的结点（起始结点），双击结点面板中要连接入数据流的结点（目标结点），这样便将数据流中的结点与新结点相连接了；



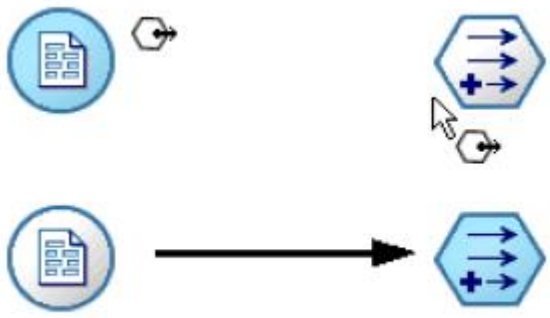
② 通过鼠标滑轮连接



在工作区内选择两个待连接的结点，用左键选中连接的起始结点，按住鼠标滑轮将其拖曳到目标结点放开，连接便自动生成。（如果鼠标没有滑轮也选用 alt 键代替）

③ 手动连接

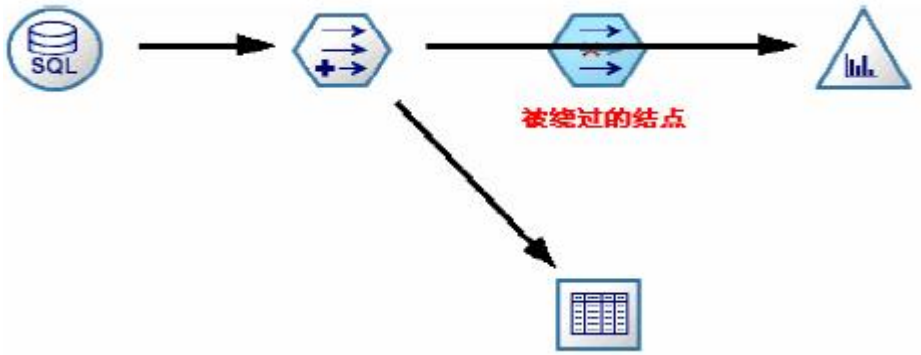
右键单击待连接的起始结点，从弹出的菜单栏中选择 Connect。选中 Connect 后鼠标和起始结点都出现了连接的标记，用鼠标单击数据流程区内要连接的目标结点，连接便生成。



注意：① 第一种连接方法是将选项面板中的结点与数据流相连接，后两种方法是将已

在数据流程区中的结点加入到数据流中

② 数据读取结点（如 SPSS File）不能有前向结点，即在连接时它只能作为起始结点而不能作为目标结点。



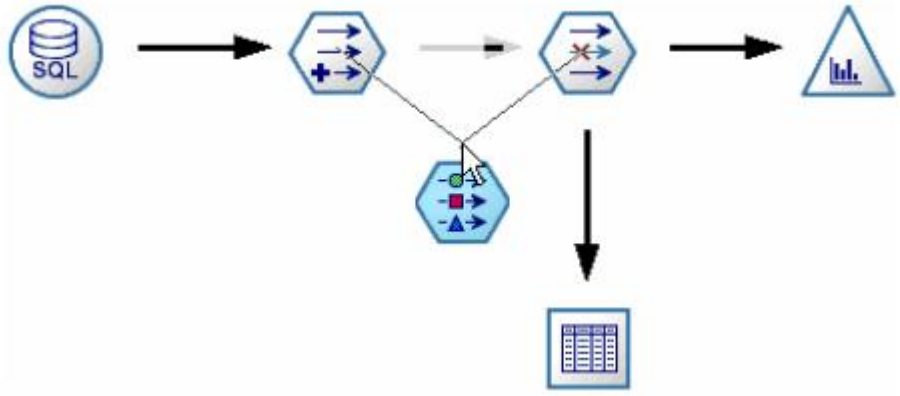
4) 绕过数据流中的结点

当我们暂时不需要数据流中的某个结点时我们可以绕过该结点。在绕过它时，如果该结点既有输入结点又有输出结点那么它的输入结点和输出结点便直接相连；如果该结点没有输出结点，那么绕过该结点时与这个结点相连的所有连接便被取消。

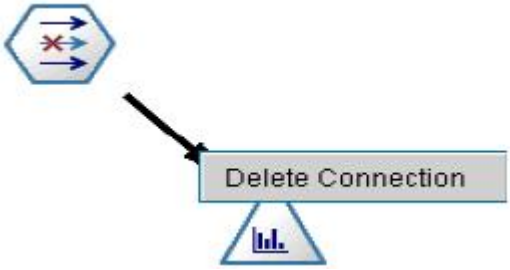
方法：用鼠标滑轮双击需要绕过的结点或者选择按住 alt 键，通过用鼠标左键双击该结点来完成。

5) 将结点加入已存在的连接中

当我们需要在两个已连接的结点中再加入一个结点时，我们可以采用这种方法将原来的连接变成两个新的连接。



方法：用鼠标滑轮单击欲插入新结点的两结点间的连线，按住它并把他拖到新结点时放手，新的连接便生成。（在鼠标没有滑轮时亦可用 alt 键代替）



6) 删除连接

当某个连接不再需要时，我们可以通过以下三种方法将它删除：

① 选择待删除的连接，单击右键，从弹出菜单中选择 Delete Connection；

② 选择待删除连接的结点，按 F3 键，删除了所有连接到该结点上的连接；

③ 选择待删除连接的结点，从主菜单中选择 Edit Node Disconnect。

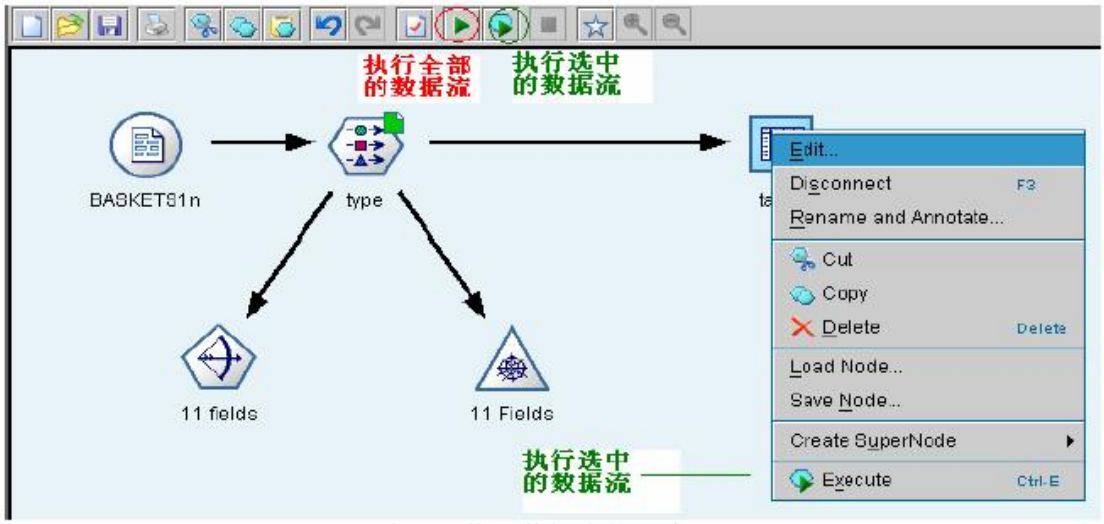
7) 数据流的执行

数据流结构构建好后要通过执行数据流数据才能从读入开始流向各个数据结点。执行数据流的方法有以下三种：

① 选择菜单栏中的按钮，数据流区域内的所有数据流将被执行；

② 先选择要输出的数据流，再选择菜单栏中的按钮，被选的数据流将被执行；

③ 选择要执行的数据流中的输出结点，单击鼠标右键，在弹出的菜单栏中选择 Execute 选项，执行被选中的数据流。



六、实验报告要求

1、能够将软件操作步骤在报告中凸现出来；

2、结果的分析要合理、准确。

七、思考题

1、数据选择在整个数据挖掘阶段的作用；

2、应用 Spss-Clementine 数据挖掘软件创建一个数据流；

3、了解相应的 SAS 数据挖掘软件的基本操作流程。

八、注意事项

1、数据流创建过程中的结点选择；

2、软件操作应当正确且熟练掌握。

实验三　数据预处理

一、实验目的

1、了解缺失值的类型。

2、解释数据挖掘过程的预处理的优点

3、掌握数据集缺失值的处理方法。

4、掌握数据挖掘缺失之处理的 Spss-Clementine 软件实现。

二、实验内容

（一）、缺失值概述

（二）、处理缺失值

1、处理带有缺失值的记录

2、处理带有缺失值的字段

3、归因或填充缺失值

4、用于缺失值的 CLEM 函数

三、实验仪器设备和材料清单

1、计算机；

2、Spss-Clementine 数据挖掘软件。

四、实验要求

教师采用课堂讲授，配合上机练习巩固所学内容的教学方法，要求学生重点掌握数据预处理的用途，能正确解释软件处理的结果，尤其是样本信息的解释；同时要求学生阅读一定数量的文献资料，加深理解如何对数据集进行合理的数据预处理。

五、实验过程

（一）缺失值概述

在数据挖掘的数据准备阶段，通常需要替换数据中的缺失值。缺失值是数据集中未知、未收集或输入不正确的值。通常，这些值不可用于字段中。例如，字段性别应包含值M和F。如果在该字段中发现值Y 或Z ，则完全可以确定此值无效，并且应将其解释为空值。同样地，年龄字段出现负值也毫无意义，应将其解释为空值。此类明显错误通常是由于问卷过程中人为输入或保留字段为空以示拒绝回答造成的。有时候，您可能会进一步检查这些空白字段，以弄清拒绝提供本人年龄等行为是否会影响具体预测结果。

某些建模技术在处理缺失值方面具有明显的优势。例如，GRI、C5.0 和Apriori 可以很好地处理在类型结点中明确声明为“missing”的值。其它建模技术在处理缺失值时比较麻烦，并且需要较长的培训时间，且生成的模型不够精确。

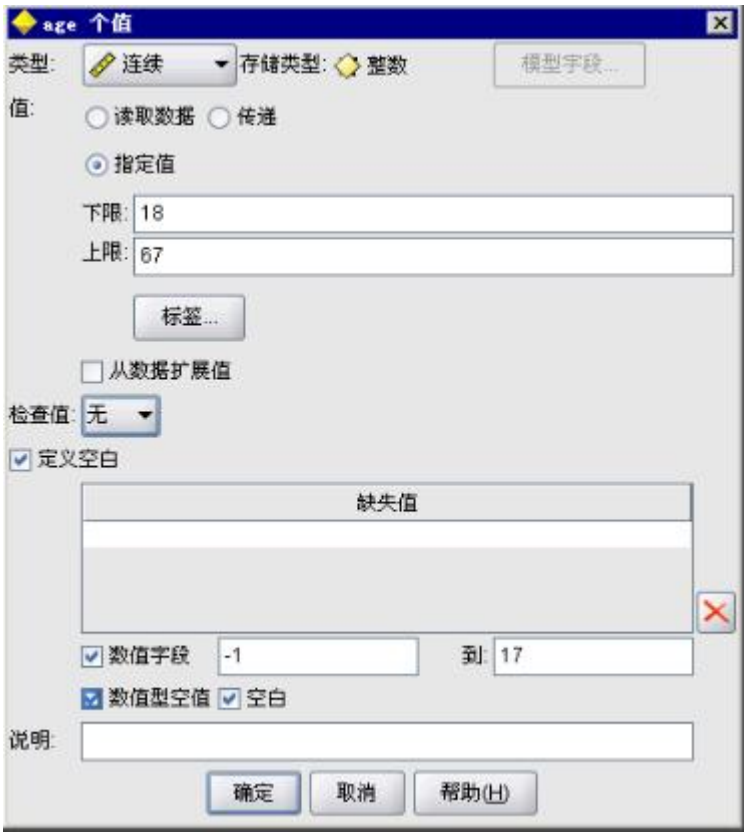
1．Clementine 可识别的缺失值类型有以下几种：

1)Null 值或系统缺失值。这两种类型是数据库或源文件中留空、并且尚未在源结点或类型结点中专门定义为“缺失”的非字符串值。系统缺失值在 Clementine 中显示为$null$。请注意，空字符串在 Clementine 中不被视为 Null 值，但它们可能会被某些数据库视为 Null 值（请参阅下面的内容）。

2)空字符串和空白。Clementine 不将空字符串和空白（带有不可见字符的字符串）视为 Null 值。对于大多数用途，空字符串都视为相当于空白。例如，如果您选择在源结点或类型结点中将空白视为空值的选项，则此设置也应用于空字符串。

3)空值或用户定义的缺失值。这些是在源结点或类型结点中被明确定义为缺失的值（如 unknown、99 或－1）。您还可以将 Null 值和空白视为空值，这样将使得它们被标记为进行特殊处理并排除在大多数计算之外。例如，您可以使用@BLANK 函数将这些值以及其他类型的缺失值处理为空值。

2．为范围变量指定缺失值



1)读取混合数据。注意，当您读取以数字形式（整数、实数、时间、时间戳或日期）存储的字段时，所有非数值型字段都将设置为 Null 值或系统缺失值。这是因为 Clementine 与其它应用程序不同，它不允许使用混合存储类型的字段。为避免发生这种情况，可以根据需要更改源结点或外部应用程序中的字段存储类型，以字符串的形式读取包含混合数据的字段。

2)从 Oracle 中读取字符型空值。在 Oracle 数据库中进行值的读写时，要注意，和 Clementine 及大多数其他数据库不一样的是，Oracle 将字符型空值等同于 Null 值对待并存储。这表示同样的数据从 Oracle 数据库中提取和从文件或其他数据库中提取其表现可能有所不同，可能会返回不同的结果。

（二）处理缺失值

您应根据自己所从事的业务或经营领域常识来确定如何处理缺失值。为了减少培训时间和提高精确度，可能需要除去数据集中的空值。从另一方面讲，空值的出现还可能会带来新的业务机会或其它灵感。选择最佳方法时，应考虑数据的以下几个方面：

1)数据集的大小

2)包含空值的字段数

3)缺失信息量

通常有两种方法可供选择：

1)可以排除带有缺失值的字段或记录

2)可以使用各种方法归因、替换或强制缺失值

使用数据审核结点可以在很大程度上实现上述两种方法的自动化。例如，您可以生成一个过滤结点，以排除带有大量缺失值且影响建模的字段，然后生成一个超结点归因全部或部分字段保留的缺失值。您可以在此引入实际审核功能，从而不仅可以访问数据的当前状态，还可以基于评估结果采取操作。

1．处理带有缺失值的记录

如果大部分缺失值都集中在少量记录中，您只需排除这些记录。例如，银行通常会保存详细而完整的贷款客户记录。但是，如果银行在审批内部职员的贷款时管制不严，则所收集的员工贷款数据可能会存在空白字段。此种情况下，有两种方法可以处理缺失值：

1) 可以使用选择结点删除员工记录。

2) 如果数据集较大，可以放弃所有带有空值的记录。

2．处理带有缺失值的字段

如果大部分缺失值都集中在少量字段中，您可以通过字段而不是记录查找这些缺失值。此方法还允许您先检验特定字段相对于建模的重要性，然后确定如何处理缺失值。如果某个字段对于建模的重要性不大，则无论它有多少缺失值，都可不必保留此字段。

例如，某市场调查公司可能会从包含 50 个问题的普通问卷中收集数据。很多人拒绝提供年龄和政治派别信息。此种情况下，年龄和政治派别就会有很多缺失值。

1)字段类型。确定要采用的方法时，您还应考虑带有缺失值的字段的类型。

2)数值字段。对于数值字段类型（如范围），您应在构建模型前清除所有非数字值，因为如果数值字段中包含空值，很多模型将无效。

3)分类字段。对于分类字段（如集合和标志），虽然不必更改缺失值，但更改后可以提高模型的精度。例如，使用性别字段的模型即使含有无意义值（如 Y 和 Z）也仍然有效，不过如果删除除 M 和 F 以外的值将提高模型的精度。

4)筛选或删除字段。要筛选带有大量缺失值的字段，您可以采用以下几种方法：

① 使用数据审核结点根据质量过滤字段。

② 您可以使用特征选择结点来筛选缺失值超过指定百分比的字段，并根据相对于特定目标的重要性来对字段进行排序。

③ 除删除字段以外，还可以使用类型结点将字段方向设置为无。此操作可将字段保留在数据集中，但不会对其进行建模操作。

3．归因或填充缺失值

在仅有几个缺失值的情况下，可以用插入值替换空值。可以在数据审核报告中实现上述操作，在此报告中您可以为特定字段指定相应选项，然后生成一个超结点采用多种方法对值进行归因。这种方法最为灵活，还可以指定在单个结点中处理大量字段。下列方法可用于输入缺失值：

1)固定。替换为固定值（可以字段平均值、范围中间值，或者您指定的常数）。

2)随机。替换为基于正态分布或均匀分布产生的随机值。

3)表达式。用于指定定制表达式。例如，您可以使用设置全局量结点创建的全局变量替换值。

4)算法。基于 C&RT 算法替换为模型预测的值。对于使用此方法输入的每个字段，都会有一个单独的 C&RT 模型，还有一个填充结点会使用该模型预测的值替换空白值和 Null 值。然后使用过滤结点删除该模型生成的预测字段。

如果还要为特定字段强制赋值，则可以使用类型结点来确保字段类型仅包含合法值，然后将需要替换空值字段的检查列设置为强制。

4．用于缺失值的 CLEM 函数

有多个 CLEM 函数可用于处理缺失值。选择结点和填充结点中经常会用以下函数来放弃或填充缺失值：

count nulls (LIST)

@BLANK (FIELD)

@NULL (FIELD)

undef

@ 函数可以与@FIELD 函数一起使用来识别一个或多个字段中是否存在空值或非 Null 值。当出现空值或非 Null 值时，一般会对此类字段进行标记，也可以用替换值填充或者在各种其它操作中使用此类字段。

如下所示，您可以计算字段列表中的非 Null 值的数量：

count_nulls(['cardtenure' 'card2tenure' 'card3tenure'])

如果要使用接受输入类型的字段列表的函数，则可以使用特定的@FIELDS_BETWEEN 和@FIELDS_MATCHING 函数，如下例所示：

count_nulls(@FIELDS_MATCHING ('card*'))

可以使用 undef 函数来填充带有系统缺失值的字段，系统缺失值显示为$null$。例如，替换数值时可以使用条件语句，如：

if not(Age > 17) or not(Age < 66) then undef else Age endif

此操作将用系统缺失值来替换所有不在该范围内的值，系统缺失值显示为$null$。借助 not() 函数，您可以获取所有其他数值，包括任何负值。

六、实验报告要求

1、能够将软件操作步骤在报告中凸现出来；

2、结果的分析要合理、准确。

七、思考题

1、为什么数据预处理是整个数据挖掘过程的重要阶段；

2、已知一本带有缺失值的思维样本：

X_1={0,1,1,2}

X_2={2,1,-,1}

X_3={1,-,-,0}

X_4={-,2,1,-}

如果所有属性的定义域是{0,1,2}，在缺失值被认为是“无关紧要的值”并且都被所给定义域的所有可行值替换的情况下，“人工”样本的数量是多少；

3、了解相应的 SAS 数据挖掘软件的基本操作流程。

八、注意事项

1、数据预处理过程中的方法的正确选择；

2、软件操作应当正确且熟练掌握。

实验四　关联规则挖掘

一、实验目的

1、解释关联规则技术的建模特性。

2、分析大型数据库的基本特性。

3、描述 Apriori 算法，并通过示例来解释算法的所有步骤。

4、通过案例了解关联分析的流程。

二、实验内容

（一）、购物篮分析

（二）、APRIORI 算法

（三）、案例分析

1、读入数据

2、关联分析

三、实验仪器设备和材料清单

1、计算机；

2、Spss-Clementine 数据挖掘软件。

四、实验要求

教师采用课堂讲授，配合上机练习巩固所学内容的教学方法，要求学生重点掌握关联规则挖掘方法的用途，能正确解释软件处理的结果，尤其是样本信息的解释；同时要求学生阅读一定数量的文献资料，加深理解关联规则挖掘方法在实际生活中的应用。

五、实验过程

（一）购物篮分析

1．基本概念：

设 $I=\{i_1,i_2,\cdots i_m\}$是项的集合，DB 为事务集合，其中每个事务 T 是项的集合，且有 $T\subseteq I$。每一个事务有一个标识符，称作 TID。 设 X 为一个项集，当且仅当 $X\subseteq T$ 时，即 T 包含 X。关联规则是形如 $X\Rightarrow Y$ 的蕴涵式，其中 $X\subseteq I$，且 $X \mathbf{I} Y=\varnothing$ 。规则 $X\Rightarrow Y$ 在事务集 DB 中成立，具有支持度 s，其中 s 是 DB 中事务包含 X 和 Y 两者的百分比。规则 $X\Rightarrow Y$ 在事务集 DB 中具有置信度 c，如果 DB 中包含 X 的事务同时也包含 Y 的百分比是 c。

支持度是概率 $P(X \mathbf{U} Y)$ 。

置信度是概率 $P(Y \mid X)$ 。

置信度可以表示规则的可信性，支持度表示模式在规则中出现的频率。具有高置信度和强支持度的规则被称为强规则。

2．挖掘关联规则的问题可以分两个阶段：

1)发掘大项集，也就是事务支持度 s 大于预先给定的最小阈值的项的集合。

2)使用大项集来产生数据库中置信度 c 大于预先给定的最小阈值的关联规则。

（二）APRIORI 算法

Apriori 算法是解决这个问题的常用方法。

Apriori 算法利用几次迭代来计算数据库中的频繁项集。第 i 次迭代计算出所有频繁 i 项集(包含 i 个元素的项集)。每一次迭代有两个步骤：产生候选集；计算和选择候选集。

在第一次迭代中，产生的候选集包含所有 1-项集，并计算其支持度 s，s 大于阈值的 1-项集被选为频繁 1-项集。

第二次迭代时，Apriori 算法首先去除非频繁 1-项集，在频繁 1-项集的基础上进行产生频繁 2-项集。原理是：如果一个项集是频繁，那么它的所有子集也是频繁的。

例如，以表 4-1 中的数据为例。假设 $S_{\min}$ =50%。

表 4-1

数据库 DB:

TID	项
001	A C D
002	B C E
003	A B C E
004	B E

在第一次迭代的第一步中，所有单项集都作为候选集，产生一个候选集列表。在下一步中，计算每一项的支持度，然后在 $S_{\min}$ 的基础上选择频繁项集。图 4-1 中给出第一次迭代的结果。

1-项集 C_1
{A}
{C}
{D}
{B}
{E}

a) 生成阶段

1-项集	计数	S[%]
{A}	2	50
{C}	3	75
{D}	1	25
{B}	3	75
{E}	3	75

b1) 计算阶段

大 1-项集 L_1	计数	S[%]
{A}	2	50
{C}	3	75
{B}	3	75
{E}	3	75

b2) 选择阶段

图 4-1 针对数据库 DB 的 Apriori 算法的第一次迭代

在挖掘 2-项集时，因为 2-项集的任何子集都是频繁项集，所以 Apriori 算法使用 L1*L1 来产生候选集。*运算通常定义为：

$$L_1 * L_1 = \{X \cup Y \text{ 其中 } X,Y \in L_k, |X \cap Y| = k+1\}$$

注:$|X \cap Y|=k+1$ 即 X 和 Y 合取容量为 k+1

当 k=1 时，因此，C2 包含在第二次迭代中作为候选集由运算 $|L_1| \cdot |L_1 - 1|/2$ 所产生的 2-项集。

本例中为：$4 \cdot 3/2=6$。用该列表来扫描 DB，计算每一个候选集的 s，并与 $S_{\min}$ 比较 2-项集 L2。图 4-2 给出了所有这些步骤和第二次迭代的结果。

2-项集 C_2
{A，B}
{A，C}
{A，E}
{B，C}
{B，E}
{C，E}

a) 生成阶段

2-项集	计数	S[%]
{A，B}	1	25
{A，C}	2	50
{A，E}	1	25
{B，C}	2	50
{B，E}	3	75
{C，E}	2	50

b1) 计算阶段

大 2-项集 L_2	计数	S[%]
{A，C}	2	50
{B，C}	2	50
{B，E}	3	75
{C，E}	2	50

b2) 选择阶段

图 4-2　针对数据库 DB 的 Apriori 算法的第二次迭代

候选集 C3 运用 $L_2 * L_2$ 来产生，运算结果得到{A,B,C},{A,C,E},{B,C,E}，但只有{B,C,E}的所有子集是频繁项集，成为候选的 3-项集。然后扫描 DB，并且挖掘出频繁 3-项集，见图 4-3 所示。

3-项集 C_3
{B，C，E}

a) 生成阶段

3-项集	计数	S[%]
{B，C，E}	2	50

b1) 计算阶段

大 3-项集 L_3	计数	S[%]
{B，C，E}	2	50

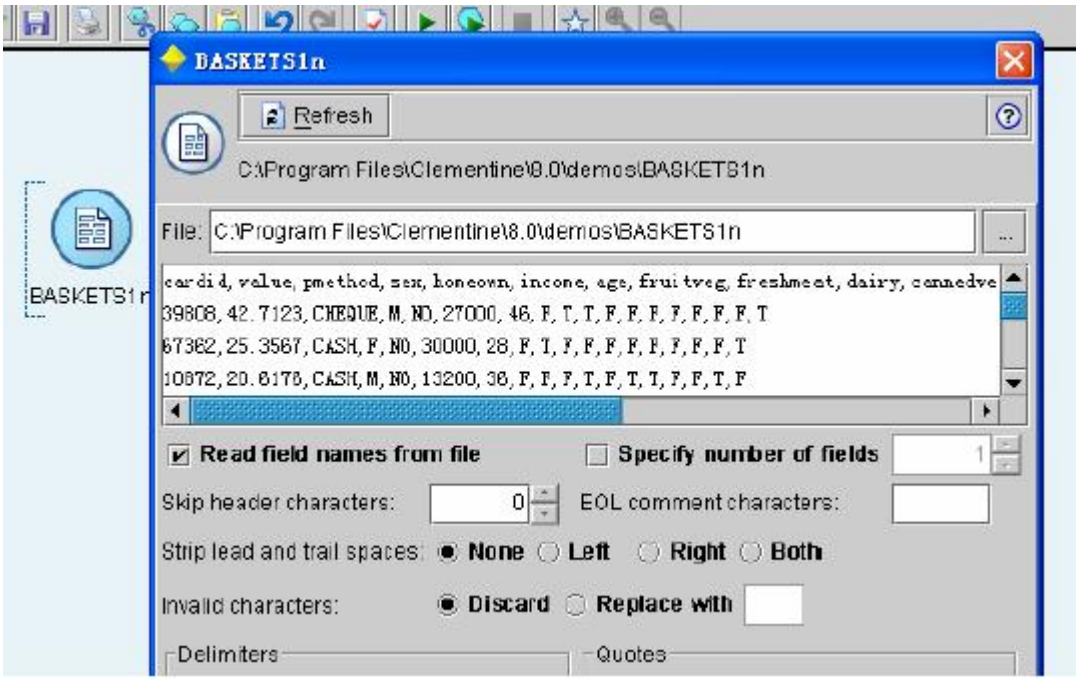
b2) 选择阶段

图 4-3　针对数据库的 DB 的 Apriori 算法的第三次迭代

因为本例的 L_3 无法产生候选的 4-项集，所以算法停止迭代过程。

该算法不仅计算所有频繁集的 s，也计算那些没有被删除的非频繁候选集的 s。所有非频繁但被算法计算 s 的候选项集的集合被称为负边界。因此，如果项集非频繁的，但它的子集都是频繁的，那么它就在负边界之中。

在本例中，负边界由项集{D},{A,B},{A,E} 组成。负边界在一些 Apriori 的改进算法中更为重要，例如生成大项集或导出负关联规则时提高了有效性。



（三）案例分析

示例 baskrule.str 是针对某商场的购物资料对数据进行分析。为了找出商品在出售时是否存在某种联系，我们将使用关联分析方法；为了得到购买某种商品的顾客特征，我们将采用决策树方法对顾客分类。

1．读入数据

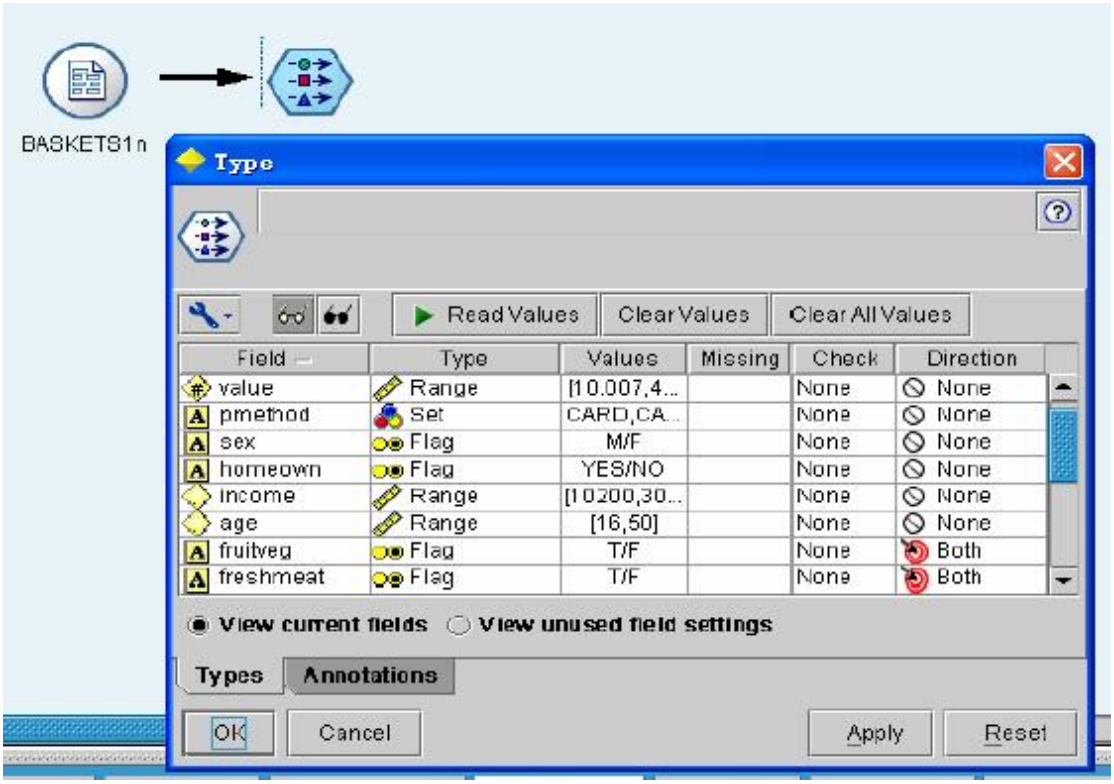
该模型的数据文件存储为 BASKETS1n，我们选择 Source 栏的 Var. File（自由格式文本文件）结点作为数据读入结点，双击该结点进行属性设置。

2．关联分析

从数据源读入数据后我们需要根据要进行的分析对字段进行设置。关联分析是分析多个量之间的关系，所以需要将进行分析的字段既设置为模型的输入又设置为模型的输出，对字段的设置可以通过 Type 结点进行。

1)为数据设置字段格式

在数据流程区内选中已存在的 Var. File 结点，双击 File Ops 栏中的 Type 结点，将 Type 结点加入到数据流中。由于我们的分析是对商品进行，与顾客的个人信息无关，所以在 Type 中将顾客个人信息字段的 Direction 设为 none，其他商品字段的 Direction 设为 Both。同时我们也将读入字段类型和字段取值。



2）生成关联分析数据流

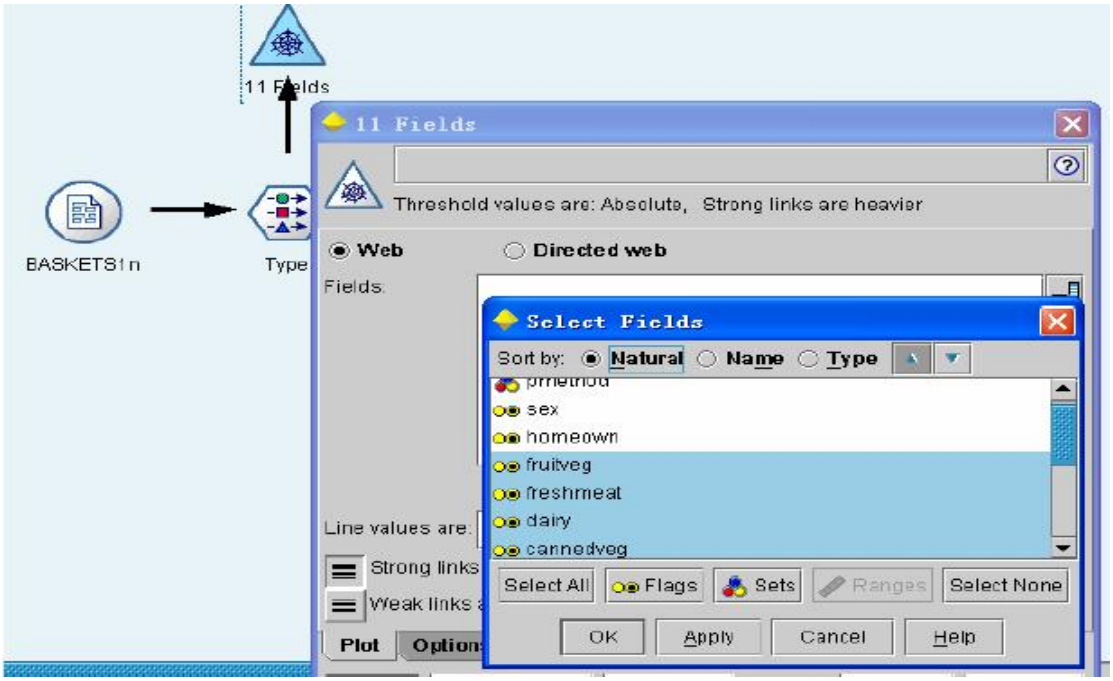
Clementine 提供了三个可以进行关联分析的模型，他们分别是 Apriori、GRI、Sequence，在这里我们选择 GRI 结点加入到数据流中。执行该数据流，它的结果将在在管理器的 Models 栏中以与模型同名的结点显示，右键选择浏览该结点，结果如下图：

结果数据表显示了各种商品间的关系，该表的每一行表明了当某种商品被购买时还有哪些产品可能被同时购买，它是居于关联分析中的支持度和可信度来分析的。

3)图形化显示各商品之间的关系

对数据进行关联分析除了利用模型外，我们还可以利用 Graphs 栏中的 Web 结点将它们之间的关系通过网状图显示。

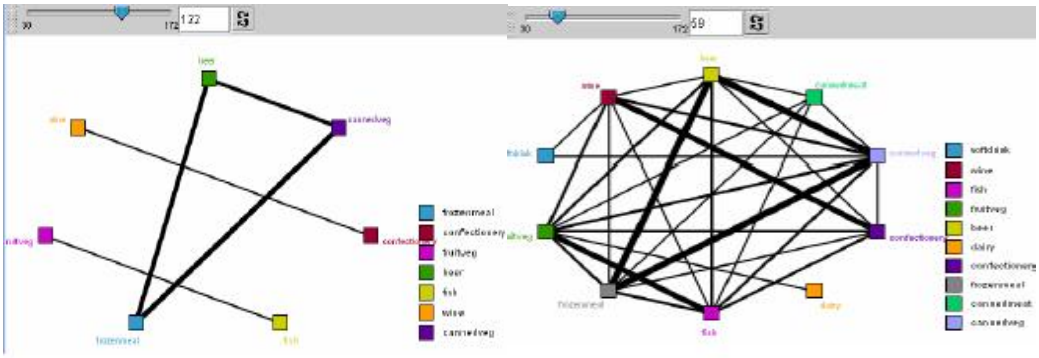
选中 Web 结点将它连接到 Type 结点上，对 Web 结点的属性设置如下图所示：



选择 Fields 栏右边的打开对话框按钮，弹出如上图所示的 Select Fields 对话框。选出将要作关联分析的项，确定后返回 Web 属性菜单。

在 plot 面板中选中“show true tag only”栏可帮我们简化输出网络。在 Web 结点的属性设置好后我们可以运行这条数据流，运行结果如下左图所示。

各色的结点代表了各种不同的商品，任两点的连线越粗表明这两点间的关系越强烈。我们还可以通过改变浮标值设置不同的显示，当浮标值越大时 web 图将显示拥有越强关系的点（如下右图所示）。



六、实验报告要求

1、能够将软件操作步骤在报告中凸现出来；

2、结果的分析要合理、准确。

七、思考题

1、Aproior 算法中，支持度和置信度参数常用的值是什么？以零售业为例解释；

2、于再事务数据库中生成大的项集相比，为什么关联规则的发现过程简单；

3、了解相应的 SAS 数据挖掘软件的基本操作流程。

八、注意事项

1、关联规则的修正与正确解释；

2、软件操作应当正确且熟练掌握。

实验五　分类：决策树

一、实验目的

1、分析解决分类问题的基于逻辑的方法的特性。

2、描述决策树和决策规则在最终分类模型中的表述之间的区别。

3、通过案例了解决策树技术的实际应用。

二、实验内容

（一）、决策树

（二）、C4.5 算法：生成一个决策树

（三）、决策树案例

1、将 Derive 结点连接到 Type 结点后

2、设置 Derive 结点的属性

3、设置字段的输入/输出方向

4、数据流的最终建立

三、实验仪器设备和材料清单

1、计算机；

2、Spss-Clementine 数据挖掘软件。

四、实验要求

教师采用课堂讲授，配合上机练习巩固所学内容的教学方法，要求学生重点掌握决策树分类方法的用途，能正确解释软件处理的结果，尤其是样本信息的解释；同时要求学生阅读一定数量的文献资料，理解决策树技术在实际生活中的应用。

五、实验过程

（一）决策树

从数据中生成分类器的一个特别有效的方法是生成一个决策树。它是一种基于逻辑的方法，通过一组输入-输出样本构建决策树的有指导学习方法。

决策树包含属性已被检验的结点，一个结点的输出分枝和该结点的所有可能的检验结果相对应。

图 5-1 是一个简单的决策树。该问题有两个属性 X，Y。所有属性值 X>1 和 Y>B 的样本属于类 2。不论属性 Y 的值是多少，值 X <1 的样本都属于类 1。

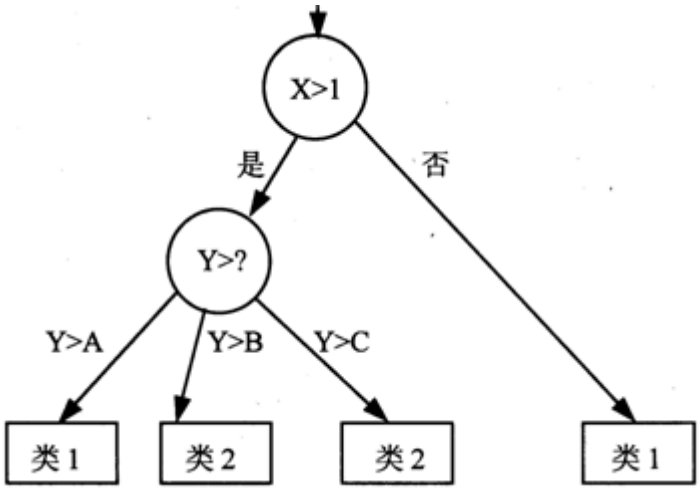


图 5-1　关于属性 X 和 Y 的检验的一个简单的决策树

对于树中的非叶结点，可以沿着分枝继续分区样本，每一个结点得到它相应的样本子集。

生成决策树的一个著名的算法是 Quinlan 的 ID3 算法，C4.5 是它改进版。

1．ID3 算法的基本思路：

1)从树的根结点处的所有训练样本开始，选取一个属性来划分这些样本。对属性的每一个值产生一个分枝。分枝属性值的相应样本子集被移到新生成的子结点上。

2)这个算法递归地应用于每个子结点，直到一个结点上的所有样本都分区到某个类中。

3)到达决策树的叶结点的每条路径表示一个分类规则。

该算法的关键性决策是对结点属性值的选择。ID3 和 C4.5 算法的属性选择的基础是基于使结点所含的信息熵最小化。

基于信息论的方法坚持对数据库中一个样本进行分类时所做检验的数量最小。ID3 的属性选择是根据一个假设，即决策树的复杂度和所给属性值表达的信息量是密切相关的。基于信息的试探法选择的是可以给出最高信息的属性，即这个属性是使样本分类的结果子树所需的信息最小。

（二）C4.5 算法：生成一个决策树

1．C4.5 算法最重要的部分是由一组训练样本生成一个初始决策树的过程。决策树可以用来对一个新样本进行分类，这种分类从该树的根结点开始，然后移动样本直至达叶结点。在每个非叶决策点处，确定该结点的属性检验结果，把注意力转移到所选择子树的根结点上。

例如：如图 5-2a 为决策树分类模型，待分类有样本如图 5-2b 所示，由决策树分类模型可得出待分类样本为类 2。(结点 A,C,F(叶结点))

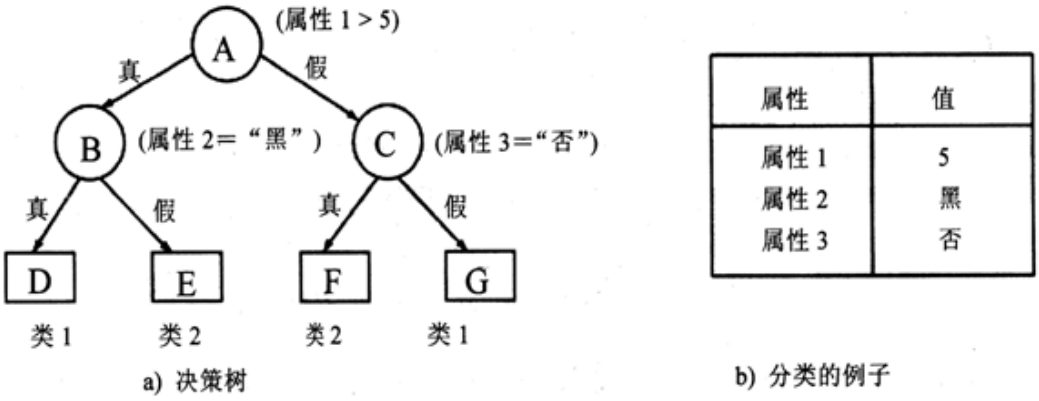


图 5-2 基于决策树模型的一个新样本的分类

C4.5 算法的构架是基于亨特的 CLS 方法，其通过一组训练样本 T 构造一个决策树。用{C1,C2,…,CK}来表示这些类，集合 T 所含的内容信息有 3 种可能性：

1)T 包含一个或更多的样本，全部属于单个的类 Cj。那么 T 的决策树是由类 Cj 标识的一个叶结点。

2)T 不包含样本。决策树也是一个叶，但和该叶关联的类由不同于 T 的信息决定，如 T 中的绝大多数类。

3)T 包含属于不同类的样本。这种情况下，是把 T 精化成朝向一个单类样本集的样本子集。根据某一属性，选择具有一个或更多互斥的输出$\{O_1,O_2,\ldots,O_n\}$的合适检验。T 被分区成子集 $T_1,T_2,\ldots,T_n$。T 的决策树包含标识检验的一个决策点和每个可能输出的一个分枝(如图 5-2a 中的 A,B 和 C 结点)

假设选择有 n 个输出(所给属性的 n 个值)的检验，把训练样本集 T 分区成子集 $T_1,T_2,\ldots,T_n$。仅有的指导信息是在 T 和它的子集 Ti 中的类分布。

如果 S 是任意样本集，设 freq(Ci,S)代表 S 中属于 Ci 的样本数量，|S|表示集合 S 中的样本数量。

2．ID3 算法的属性选择的检验方法采用增益标准，它基于信息论中熵的概念。

集合 S 的期望信息(熵)如下：

$$info(S)=-\sum_{i=1}^{k}((freg(C_i,S)/|S|)\cdot\log_2(freg(C_i,S)/|S|))$$

T 被分区之后的一个相似度标准，T 按照一个属性检验 X 的几个输出进行分区。所需信息为子集的熵的加权和:

$$info_x(T)=-\sum_{i=1}^{n}(|T_i|/|T|)\cdot info(T_i))$$

分区所对应的信息增益:

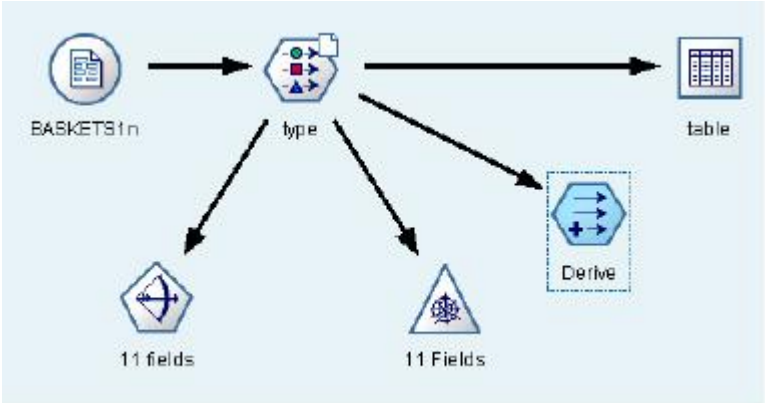
$$Gain(X)=info(T)-info_x(T)$$

上式度量了按照检验 X 进行分区的 T 所得到的信息。该增益标准选择了使 Gain(X)最大化的检验 X，即此标准选择的具有最高增益的那个属性。

（三）决策树案例

在本例中我们运用决策树对购买某样商品的客户进行分类，通过分析他的个人信息（例如年龄、收入等）判断怎样的人会购买健康食品。在用决策树建模时我们需要设置一个输出结点，模型根据样本在该结点的不同取值构造出决策树。

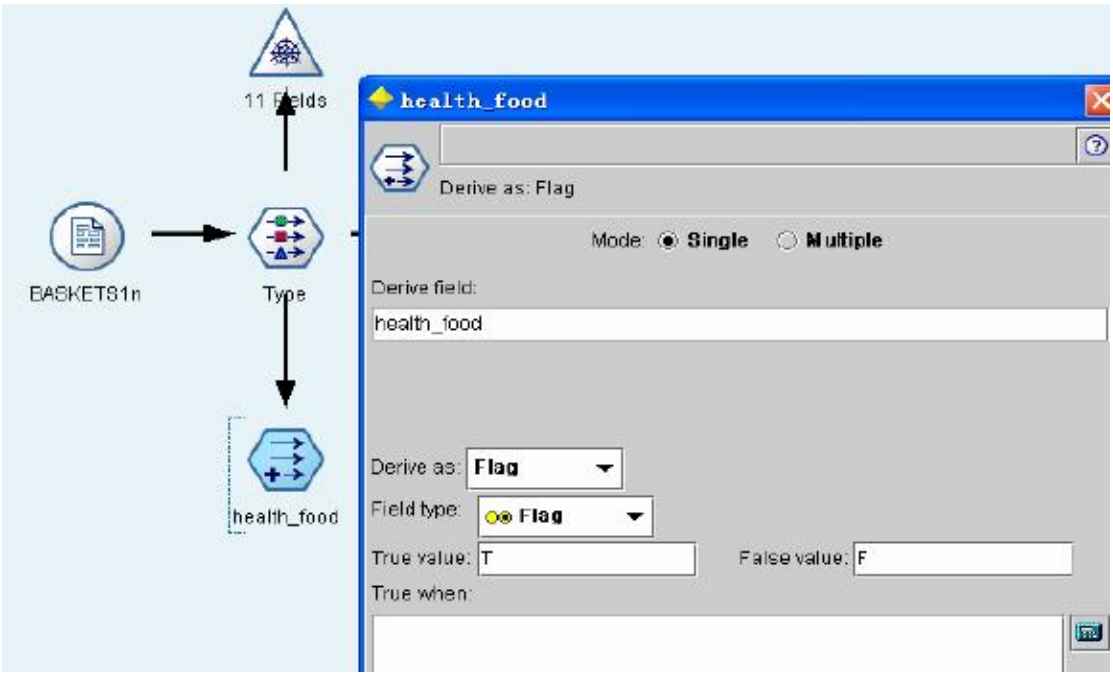
1．将 Derive 结点连接到 Type 结点后



Derive 结点在 Field OPs 栏中，可选用任何一种结点连入数据流的方法将这个结点连接；

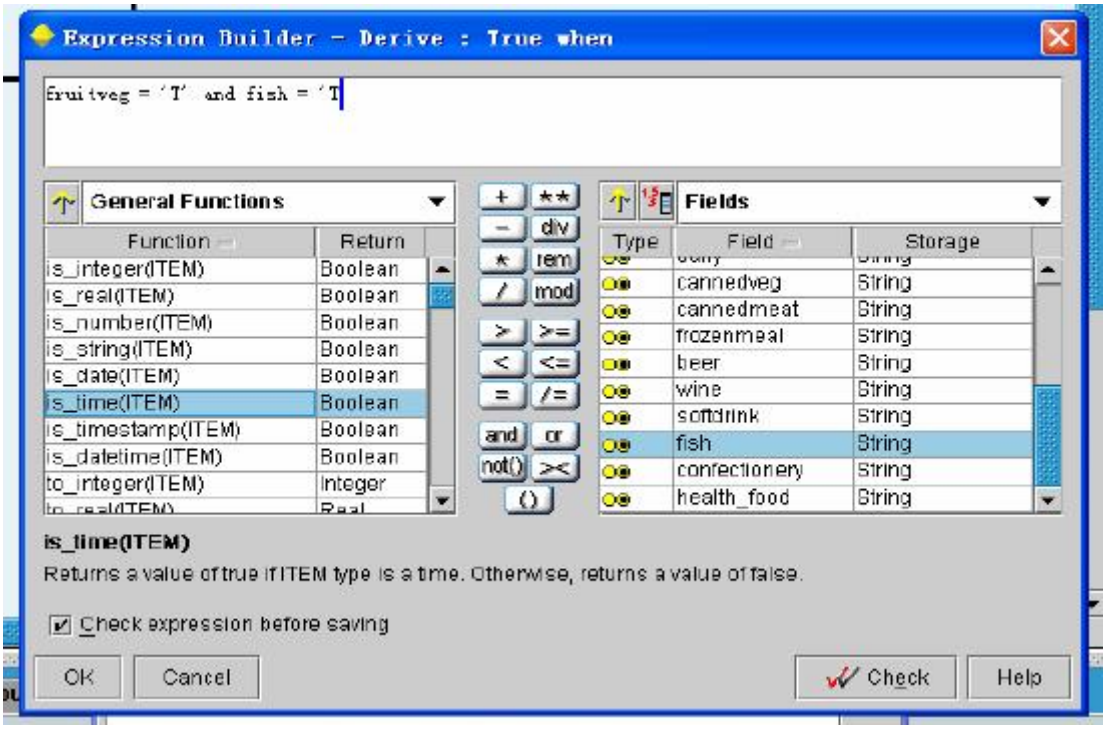
2．设置 Derive 结点的属性

双击 Derive 结点打开属性对话框，如下图所示：



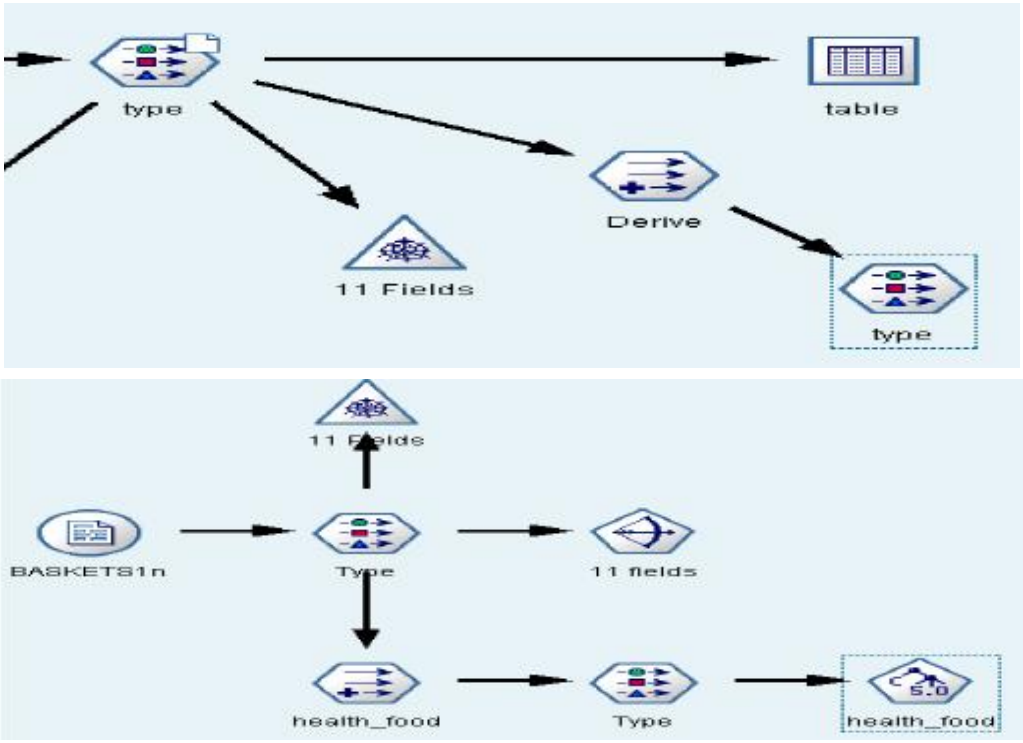
在 Drive Field 栏中将该结点命名为 health_food，在 Drive as 栏中选择 Flag，这表明新生成的 health_food 字段将存储两种数值类型的数据。在 True value 和 False value 栏中分别填写新字段的两种数据值，其中 True value 表示当条件满足时该字段的值，False value 表明当条件不满足时该字段的值。

对判断条件的设置我们可以通过单击 True when 栏右边的按钮进行。在 Expression Builder 中我们可以选择数据的任一字段，通过设计表达式建立结果为真时的条件。这里我们设置表达式为 fruitveg = 'T' and fish = 'T'，这表明当顾客购买了 fruitveg 和 fish 时该顾客便购买了健康食物。



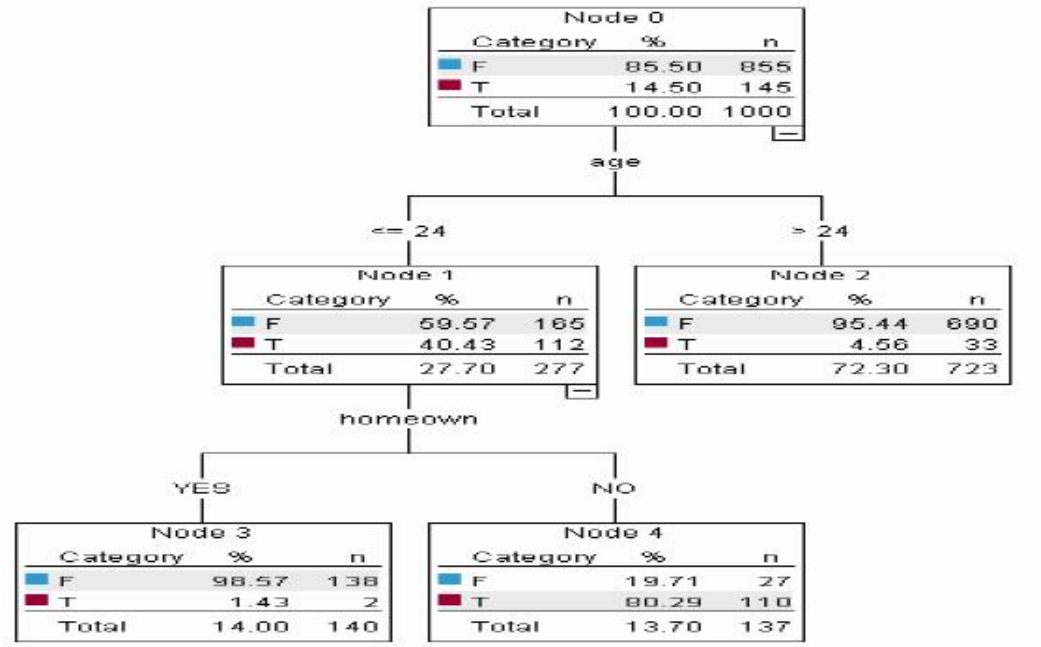
3．设置字段的输入/输出方向

要用决策树模型建模就需要在数据载入模型前定义一个输出字段，这里我们通过在 health_food 结点后添加一 Type 结点来定制字段的输入/输出方向。由于我们要分析购买健康食物的顾客特征，所以我们将 health_food 字段的 Direction 选项设置为 Out，将顾客的各个特征设置为 In，将其他商品设置为 None。



4．数据流的最终建立

在对字段定义结束后，我们将 C5.0（决策树模型）结点加入到数据流。其数据流建立如下图：



运行建立了决策树的数据流，我们可得到输出结果如上树形图所示。该树的叶结点表

明了怎样的顾客将选择健康食品，怎样的顾客将拒绝健康食品，我们也可以根据该树的将客户按是否购买健康食品进行分类

P.S.：在这个决策树分析的案例中我们用到了 Var. File、Derive、Web、GRI 和 C5.0 结点。

六、实验报告要求

1、能够将软件操作步骤在报告中凸现出来；

2、结果的分析要合理、准确。

七、思考题

1、假定有 C4.5 生成的两个决策规则：

规则 1：（X>3）$\cap$(Y≥2)$\rightarrow$Class1(9.6/0.4)

规则 2：（X>3）$\cap$(Y<2)$\rightarrow$Class2(2.4/2.0)

分析是否有可能用二项式分布的置信极限$U_{25\%}$把这两个规则化成一个；

2、现实数据挖掘应用中，最终模型包含大量的决策规则。讨论并分析应该怎样做才能简化模型的复杂性；

3、了解相应的 SAS 数据挖掘软件的基本操作流程。

八、注意事项

1、决策规则的合理修剪与正确认识；

2、软件操作应当正确且熟练掌握。

实验六　因子分析

一、实验目的

1、概述因子分析。

2、了解数据挖掘中的因子分析技术。

3、通过案例讲解掌握因子分析的实际操作流程。

二、实验内容

（一）、读入数据

（二）、设置字段属性

（三）、对数据进行因子分析

（四）显示经过因子分析后的数据表

1、为因子变量命名

2、数据输出显示

三、实验仪器设备和材料清单

1、计算机；

2、Spss-Clementine 数据挖掘软件。

四、实验要求

教师采用课堂讲授，配合上机练习巩固所学内容的教学方法，要求学生重点掌握因子分析的用途，能正确解释软件处理的结果，尤其是样本信息的解释；同时要求学生阅读一定数量的文献资料，加深理解因子分析的应用。

五、实验过程

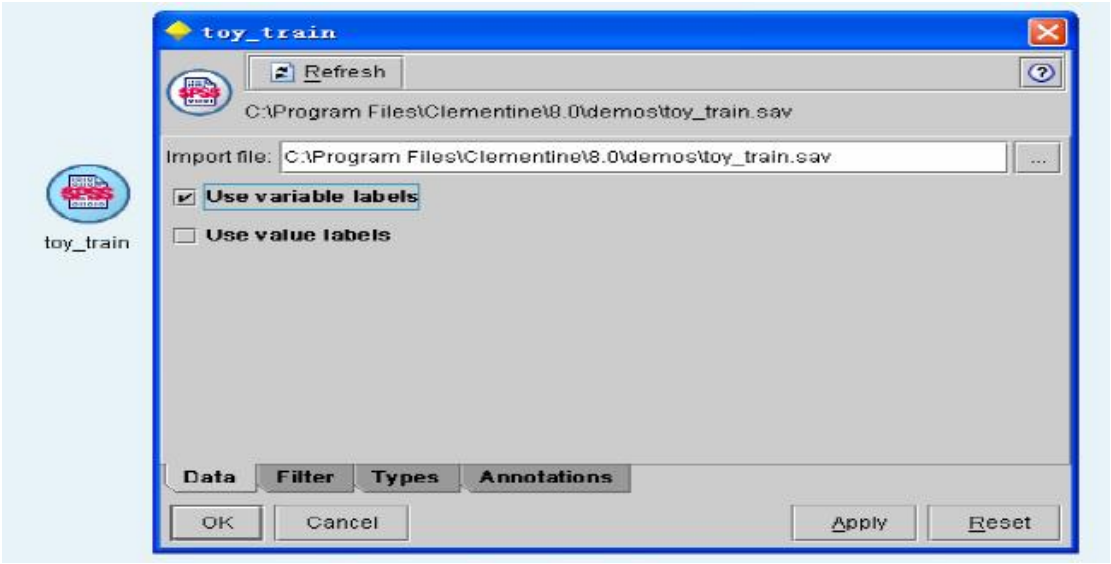
因子分析(factor. str)

示例 factor.str 是对孩童的玩具使用情况的描述，它一共有 76 个字段。过多的字段不仅增添了分析的复杂性，而且字段之间还可能存在一定的相关性，于是我们无需使用全部字段来描述样本信息。下面我们将介绍用 Clementine 进行因子分析的步骤：

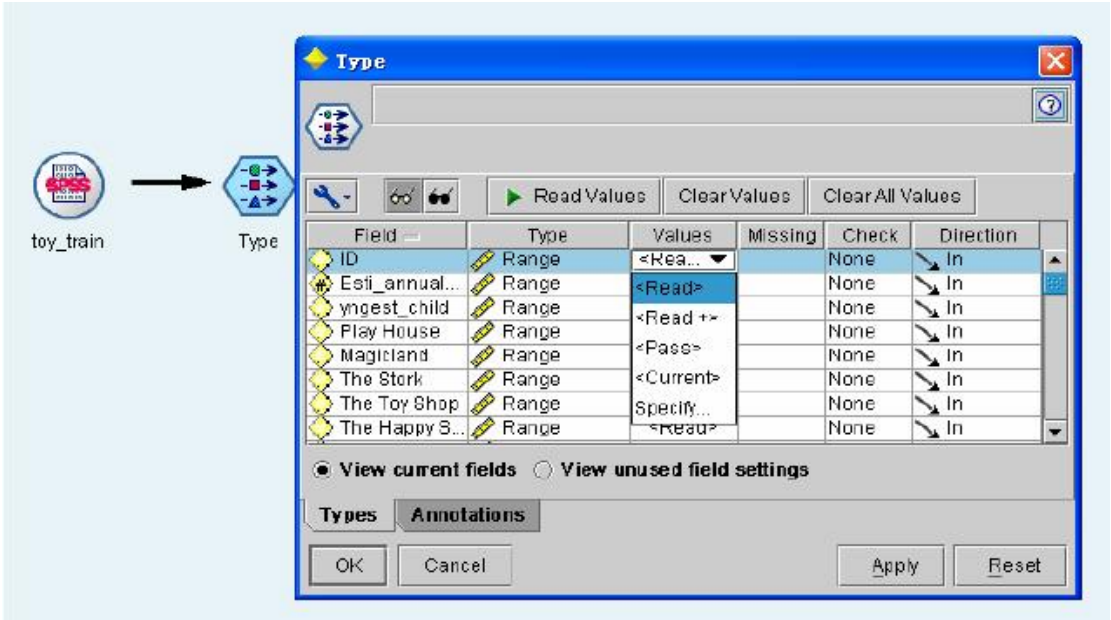
（一）读入数据

Source 栏中的结点提供了读入数据的功能，由于玩具的信息存储为 toy_train.sav，所以我们需要使用 SPSS File 结点来读入数据。双击 SPSS File 结点使之添加到数据流程区内，双击添加到数据流程区里的 SPSS File 结点，由此来设置该结点的属性。

在属性设置时，单击 Import file 栏右侧的按钮，选择要加载到数据流中进行分析的文件，这里选择 toy_train.sav。单击 Annotations 页，在 name 栏中选择 custom 选项并在其右侧的文本框中输入自定义的结点名称。这里我们按照原示例输入 toy_train。



（二）设置字段属性



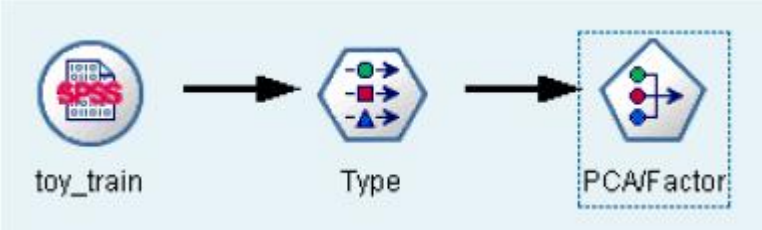
进行因子分析时我们需要了解字段间的相关性，但并不是所有字段都需要进行相关性分析，比如“序号”字段，所以需要我们将要进行因子分析的字段挑选出来。Field Ops 栏中的 Type 结点具有设置各字段数据类型、选择字段在机器学习中的输入/输出属性等功能，我们利用该结点选择要进行因子分析的字段。首先，将 Type 结点加入到数据流中，双击该

结点对其进行属性设置。

由上图可看出数据文件中所有的字段名显示在了 Field 栏中，Type 表示了每个字段的数据类型。我们不需要为每个字段设定数据类型，只需从 Values 栏中的下拉菜单中选择<Read>项，然后选择 Read Value 键，软件将自动读入数据和数据类型；Missing 栏是在数据有缺失时选择是否用 Blank 填充该字段；Check 栏选择是否判断该字段数据的合理性；而 Direction 栏在机器学习模型的建立中具有相当重要的作用，通过对它的设置我们可将字段设为输入/输出/输入且输出/非输入亦非输出四种类型。在这里我们将前 19 个字段的 Direction 设置为 none，这表明在因子分析我们不将这前 19 个字段列入考虑，从第 20 个字段起我们将以后字段的 direction 设置为 In，对这些字段进行因子分析。

（三）对数据进行因子分析

因子分析模型在 Modeling 栏中用 PCA/Factor 表示。在分析过程中模型需要有大于或等于两个的字段输入，上一步的 Type 结点中我们已经设置好了将作为模型输入的字段，这里我们将 PCA/Factor 结点连接在 Type 结点之后不修改它的属性，默认采用主成分分析方法。



在建立好这条数据流后我们便可以将它执行。右键单击 PCA/Factor 结点，在弹出的菜单栏中选择 Execute 执行命令。执行结束后，模型结果放在管理器的 Models 栏中，其标记为名称为 PCA/Factor 的黄色结点。

右键单击该结果结点，从弹出的菜单中选择 Browse 选项查看输出结果。由结果可知参与因子分析的字段被归结为了五个因子变量，其各个样本在这五个因子变量里的得分也在结果中显示。

（四）显示经过因子分析后的数据表

模型的结果结点也可以加入到数据流中对数据进行操作。我们在数据流程区内选中 Type 结点，然后双击管理器 Models 栏中的 PCA/Factor 结点，该结点便加入到数据流中。为了显示经过因子分析后的数据我们可以采用 Table 结点，该结点将数据由数据表的形式输出。

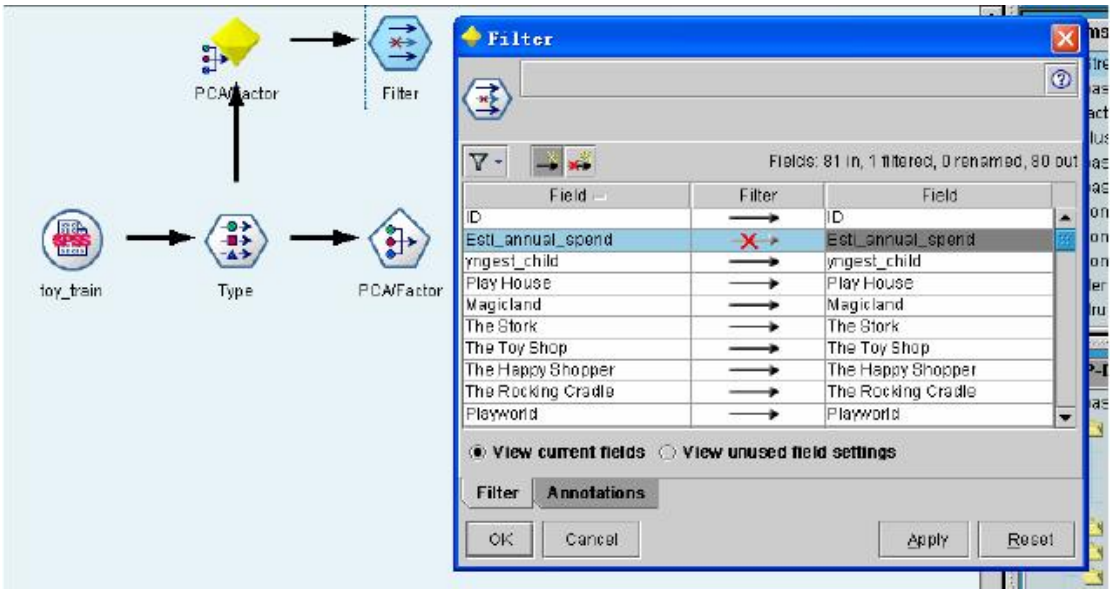
1．为因子变量命名

在将 PCA/Factor（结果）结点连接到 Table 结点之前，用户可以设置不需要显示的字段，也可以更改因子变量名，为了达到这个目的我们可以添加 Field Ops 栏中的 filter 结点。

在对 filter 结点进行属性设置时，Filter 项显示了字段的过滤与否，如果需要将某个字段过滤，只需用鼠标单击 Filter 栏中的箭头，当箭头出现红“×”时该字段便被过滤。

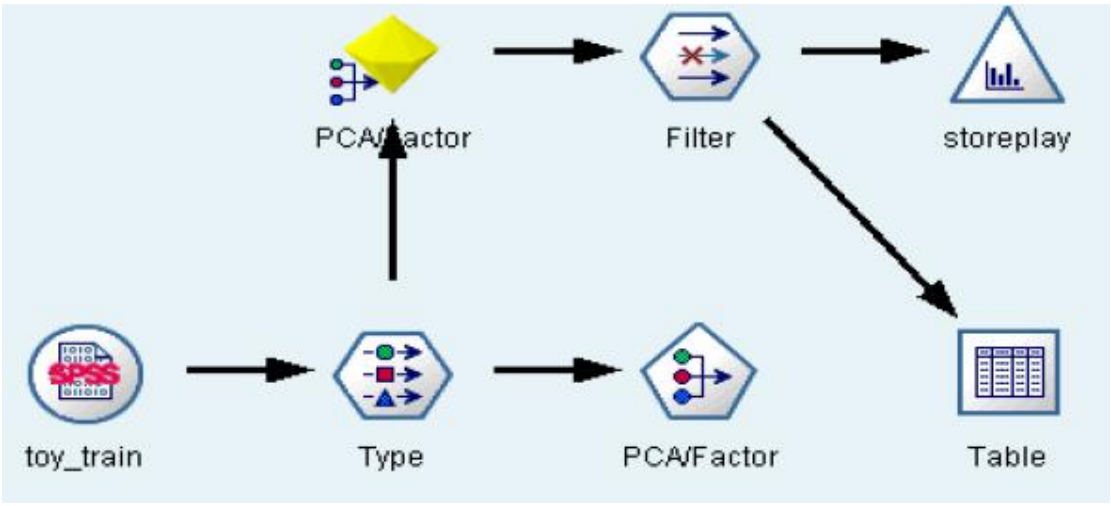
第一个 Field 栏结点表明数据在读入 filter 结点时的字段名，第二个 Field 栏表示数据经过 filter 结点后的字段名。由于因子分析生成的因子变量都由系统自动命名，用户可以通过

修改这些因子变量的第二个 Field 的值来重新设定其字段名。



2．数据输出显示

在对数据进行输出时我们选择了 Output 栏中的 Table 结点和 Graph 栏中的 Histogram 结点。这两个结点一个通过数据表的形式输出，一个通过柱状图的形式输出。对柱状图我们设置其显示 store play 字段的数据（store play 为第五个因子变量的新名）。通过“执行”按钮分别执行两条数据流，将经过因子分析后的数据显示。



P.S. ：在这个因子分析的案例中我们用到了 SPSS File、Type、Filter、Table、Histogram、PCA/Factor 结点。

六、实验报告要求

1、能够将软件操作步骤在报告中凸现出来；

2、结果的分析要合理、准确。

七、思考题

1、数据挖掘技术当中的因子分析方法同传统统计学当中的因子分析方法的区别；

2、了解相应的 SAS 数据挖掘软件的基本操作流程。

八、注意事项

1、如何合理的解释因子分析的结果；

2、软件操作应当正确且熟练掌握。

实验七　聚类分析

一、实验目的

1、了解类有不同表示法和相似度的不同量度标准。

2、用相似度的单链接或全链接度量标准实现凝聚算法。

3、推导分区聚类的 K-平均法并分析其复杂性。

4、通过案例讲解掌握聚类分析。

二、实验内容

（一）、读入数据

（二）、为数据设置字段格式

（三）、生成聚类分析数据流

（四）、图形化输出各个类的组成情况

1、将模型的结果结点连入数据流

2、设置图形输出结点

三、实验仪器设备和材料清单

1、计算机；

2、Spss-Clementine 数据挖掘软件。

四、实验要求

教师采用课堂讲授，配合上机练习巩固所学内容的教学方法，要求学生重点掌握聚类分析方法的用途，能正确解释软件处理的结果，尤其是样本信息的解释；同时要求学生阅读一定数量的文献资料，理解聚类分析方法在实际生活中的应用。

五、实验过程

Clementine 提供了多种可用于聚类分析的模型，包括 Kohonen，K-means，Two-Step 方法。

示例 Cluster.str 是对人体的健康情况进行分析，通过测量人体类胆固醇、Na、Ka 等的含量将个体归入不同类别。示例中采用了三种方法对数据进行分类，这里我们重点讨论 K-means 聚类方法。

（一）读入数据

和前两步一样，在建立数据流时首先应读入数据文件。该示例中数据文件存储为 DRUG1n，我们向数据流程区内添加 Var. File 结点读入数据。

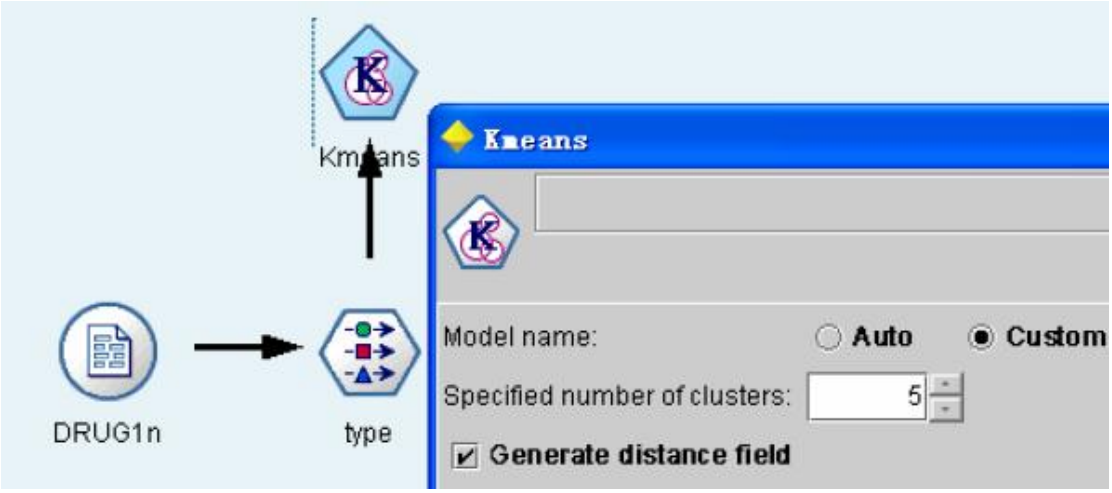
（二）为数据设置字段格式

将 Type 结点连入数据流，通过编辑该结点对数据字段进行设置。

在机器学习方法中聚类被称为无导师的学习。所谓无导师的学习是指事先并不知道数据的分类情况，就像在决策树方法中我们通过已知的某个结点值来建立模型，在聚类方法中所有参与聚类的字段在设置字段格式时其 Direction 都将被设置为 In。

（三）生成聚类分析数据流

设置好字段格式后我们将 K-means 结点加入到数据流。在编辑 K-means 结点时我们重点需要定义将要分成的类别数，这个属性在 Specified number of cluster 中设定。

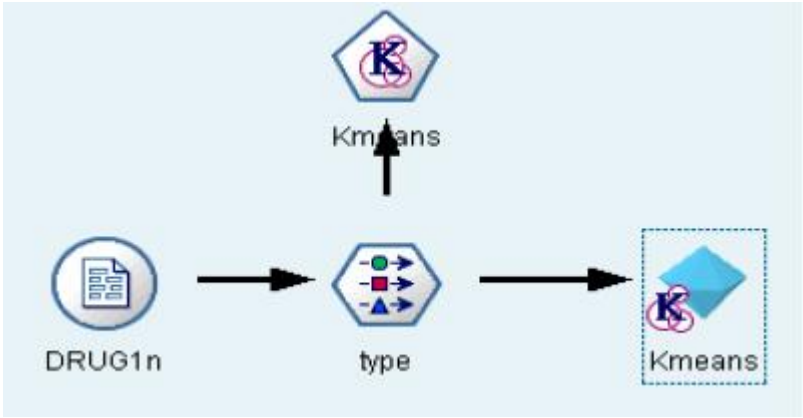


数据流建立好后，右键单击 K-means 结点选择执行该数据流。执行结果以与 K-mean 同名的结点显示在管理器的 Models 窗口中，浏览该结点我们能够得到关于分类的信息，如下图所示：

（四）图形化输出各个类的组成情况

查看各类中的情况除了浏览结果结点外，我们还可以选择用图形将结果显示出来。

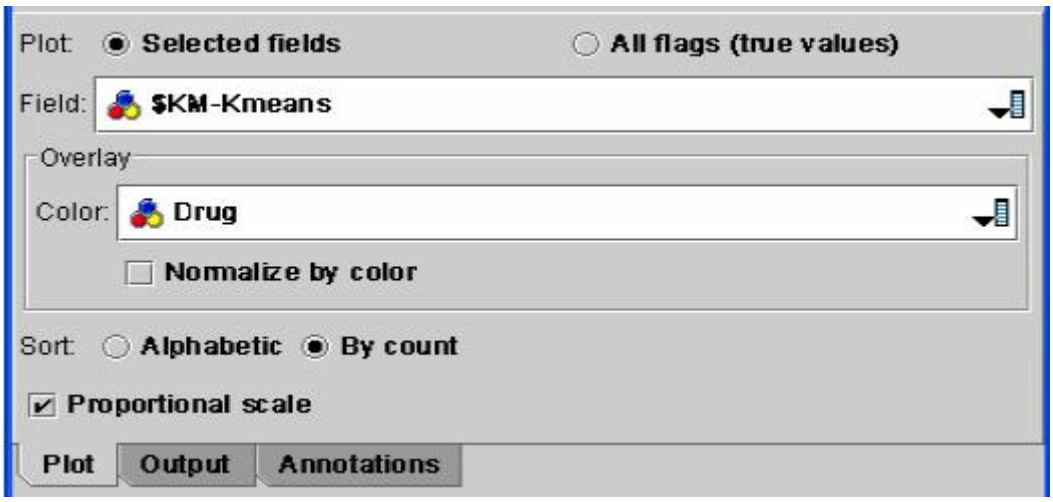
1．将模型的结果结点连入数据流



选中 Type 结点，双击 Models 窗口中的 K-means 结果结点将该结点连接到 Type 后

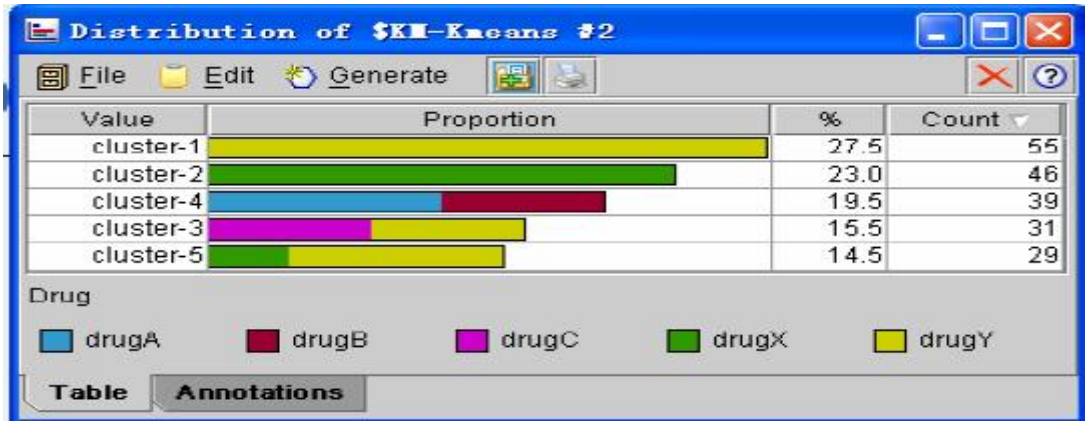
2．设置图形输出结点

选择 Graph 栏中的 Distribution 结点将它连接到 K-means 结果结点后，双击该结点对它进行设置。



在 Field 栏中选择$KM-K-means 选项，该选项保存了分类结果，即每个样本在聚类后所属的类别。Distribution 结点要求 Field 栏为非数据结点。在 Overlay 选项中我们选择 Drug 项，这是为了研究在不同的分类类别里 Drug 的各个取值的所占比例。

运行该数据流我们可得到下图,图中详细的显示了不同 Drug 类型在各个类别里的分布情况。同样道理，我们也可以对其他属性进行研究。



P.S.: 在这个聚类分析的案例中我们用到了 K-means、Distribution 结点。

六、实验报告要求

1、能够将软件操作步骤在报告中凸现出来；

2、结果的分析要合理、准确。

七、思考题

1、为什么一个聚类过程的确认是非常主观的；

2、什么增加了聚类算法的复杂度；

3、了解相应的 SAS 数据挖掘软件的基本操作流程。

八、注意事项

1、聚类分析的结果的合理的解释；

2、软件操作应当正确且熟练掌握。

实验八　神经网络

一、实验目的

1、认识神经网络的基本组成以及它们的属性和功能。

2、描述神经网络通常执行的学习任务，如模式关联、模式识别、估计、控制以及过滤

3、了解神经网络的基本流程。

4、通过案例掌握神经网络操作技术。

二、实验内容

（一）、读入数据

（二）、计算促销前后销售额的变化率

（三）、为数据设置字段格式

（四）、神经网络学习过程

（五）、为训练网络建立评估模型

1、将模型结果结点连接到数据流

2、添加字段比较预测值与实际值

3、评价模型

（六）模型预测

1、预测模型建立

2、输出规范化

3、选择促销方案

三、实验仪器设备和材料清单

1、计算机；

2、Spss-Clementine 数据挖掘软件。

四、实验要求

教师采用课堂讲授，配合上机练习巩固所学内容的教学方法，要求学生重点掌握神经网络方法的用途，能正确解释软件处理的结果，尤其是样本信息的解释；同时要求学生阅读一定数量的文献资料，理解神经网络方法在实际生活中的应用。

五、实验过程

神经网络（goodlearn.str）

神经网络是一种仿生物学技术，通过建立不同类型的神经网络可以对数据进行预存、分类等操作。

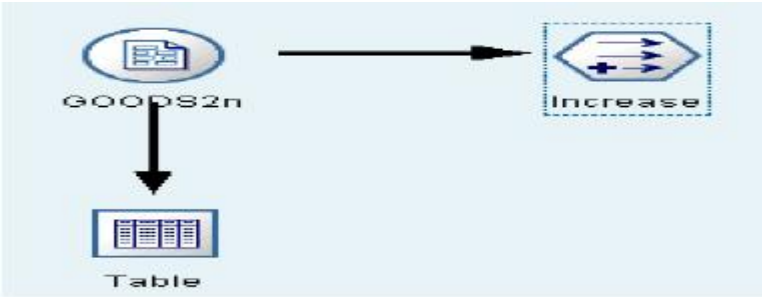
示例 goodlearn.str 通过对促销前后商品销售收入的比较，判断促销手段是否对增加商品收益有关。Clementine 提供了多种预测模型，包括 Nerual　Net、Regression 和 Logistic。这里我们用神经网络结点建模，评价该模型的优良以及对新的促销方案进行评估。

（一）读入数据

本示例的数据文件保存为 GOODS1n，我们向数据流程区添加 Var. File 结点，并将数据文件读入该结点。

（二）计算促销前后销售额的变化率

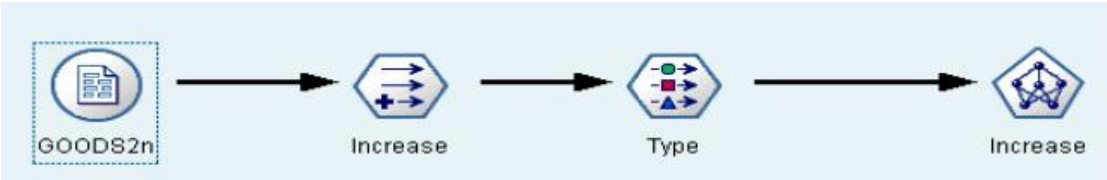
向数据流增加一个 Derive 结点，将该结点命名为 Increase。在公式栏中输入(After - Before) / Before * 100.0 以此来计算促销前后销售额的变化。



（三）为数据设置字段格式

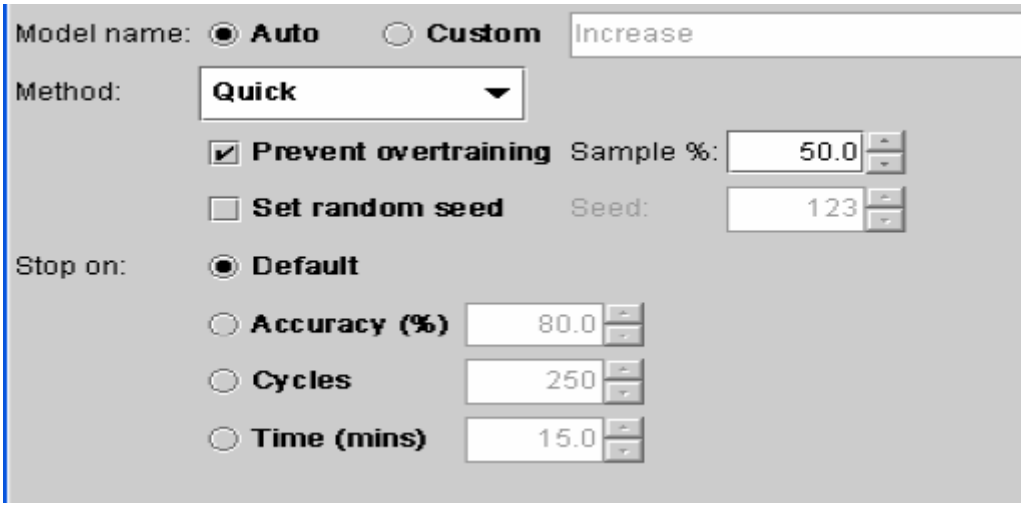
添加一个 Type 结点到数据流中。由于在制定促销方案前我们并不知道促销后商品的销售额，所以将字段 After 的 Direction 属性设置为 None；神经网络模型需要一个输出，这里我们将 Increase 字段的 Direction 设置为 Out，除此之外的其它结点全设置为 In。

（四）神经网络学习过程



在设置好各个字段的 Direction 方向后我们将 Neural Net 结点连接入数据流。

在对 Neural Net 进行设置时我们选择快速建模方法（Quick），选中 Prevent overtraining 防止过度训练。同时我们还可以根据自己的需要设置训练停止的条件。



在建立好神经网络学习模型后我们运行这条数据流，结果将在管理器的 Models 栏中显示。选择查看该结果结点，我们可以对生成的神经网络各个方面的属性有所了解。

（五）为训练网络建立评估模型

1．将模型结果结点连接到数据流

将 Increase 结果结点连接在数据流中的 Type 结点后；

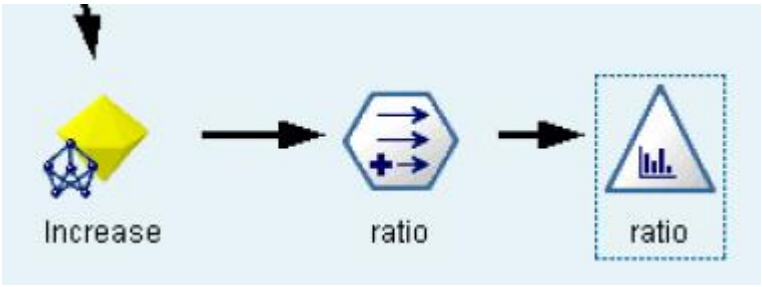
Class	Cost	Promotion	Before	After	$N-Increase
Luxury	31.2...	1467	2233...	238333	5.359
Drink	82.5...	1316	1989...	219791	9.903
Luxury	10.4...	1734	2480...	266357	6.531
Drink	40.4...	1002	2159...	235013	7.865
Drink	20.2...	1127	2890...	305659	8.335
Meat	59.3...	1884	2347...	241302	4.338
Meat	71.1...	1655	2087...	216708	3.567
Drink	62.7...	1108	1922...	204458	8.668
Drink	98.2...	1075	2342...	248692	8.574

2．添加字段比较预测值与实际值

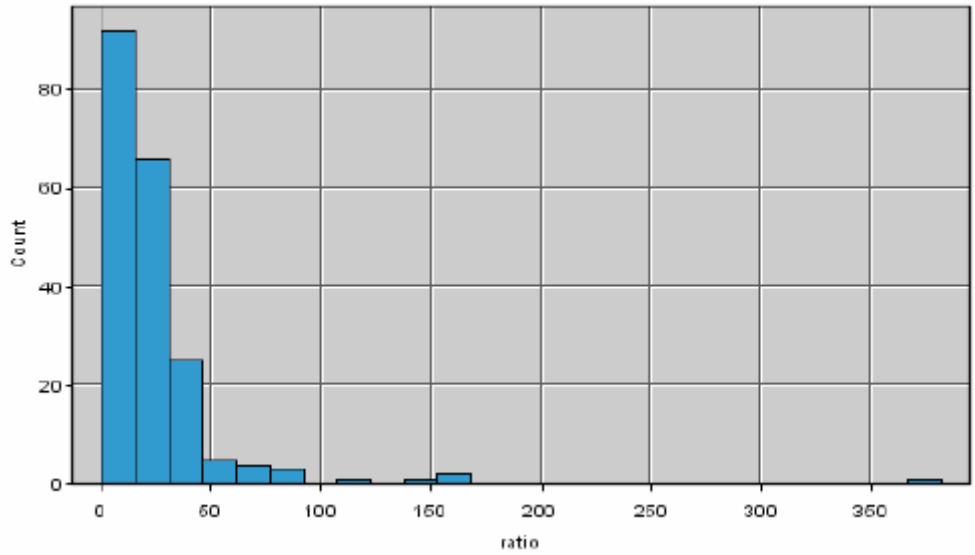
向数据流中增加 Derive 结点并将它命名为 ratio，然后将它连接到 Increase 结果结点。设置该结点属性，将增添的字段的值设置为(abs(Increase - '$N-Increase') /Increase) * 100，其中$N-Increase 是由神经网络生成的预测结果。通过该字段值的显示我们可以看出预测值与实际值之间的差异大小。

3．评价模型

可以通过观察预测值与实际值之间的差异来评价模型的优劣。从 Graph 栏中选择 histogram 结点连接到 ratio 结点。



设置该结点，使其输出显示 ratio 的值（在 field 的下拉列表中选择 ratio），输出结果如下图所示：

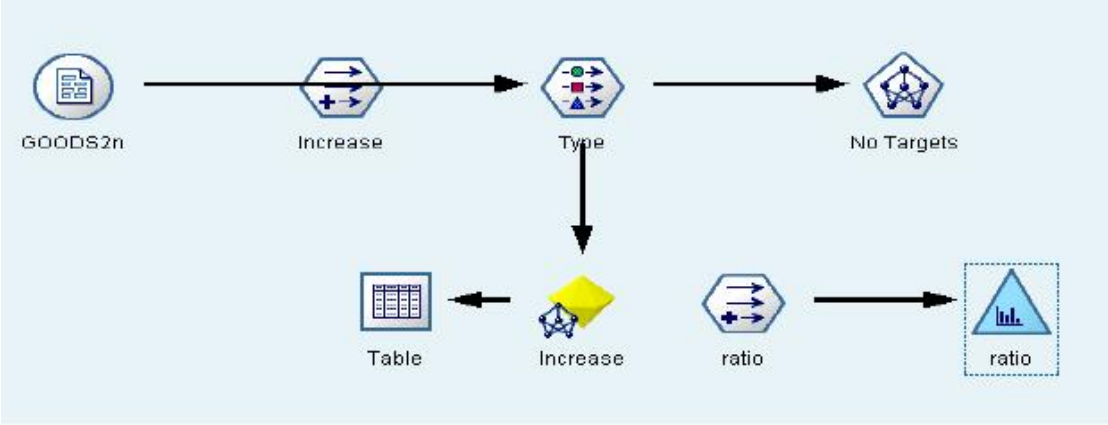


该图形的横坐标为 ratio 的值，纵坐标表示一共有多少个样本的 ratio 值落在相对应的横坐标上。从 ratio 的定义公式我们知道 ratio 越小表明预测值与实际值的差别越小，所以我们

希望更多的 ratio 值处于一个比较小的范围。因此由输出图形我们可以看出该模型达到了一定的精度。

（六）模型预测

1．预测模型建立



该模型的建立就是为了预测新样本。我们现将数据源的文件改为 GOODS2n；然后按住 alt 键双击 Increase 结点以此来绕过该结点；断开 Increase 结果结点与 Ratio 结点之间的连接，再增添一个 Table 结点观察 Increase 结果结点的输出。在 Type 结点中我们只设置字段 after 的 Direction 属性为 None，其余的都为 In。通过这种方法建立好的数据流如下图所示：

右键单击 Table 结点，选择运行数据流。运行生成的结果如下，其中$N-Increase 为预测结果：

	Class	Cost	Promotion	Before	After	$N-Increase
1	Luxury	31.2...	1467	2233...	238333	5.359
2	Drink	82.5...	1316	1989...	219791	9.903
3	Luxury	10.4...	1734	2480...	266357	6.531
4	Drink	40.4...	1002	2159...	235013	7.865
5	Drink	20.2...	1127	2890...	305659	8.335
6	Meat	59.3...	1884	2347...	241302	4.338
7	Meat	71.1...	1655	2087...	216708	3.567
8	Drink	62.7...	1108	1922...	204458	8.668
9	Drink	98.2...	1075	2342...	248692	8.574
10	Drink	34.6...	1644	1109...	121988	11.419
11	Luxury	87.4...	1105	1361...	140323	4.465
12	Drink	92.7...	1828	2091...	239858	12.096
13	Luxury	66.4...	1137	1218...	126166	4.420
14	Meat	5.810	1446	2062...	214172	2.630
15	Meat	92.9...	1260	1574...	159442	2.672

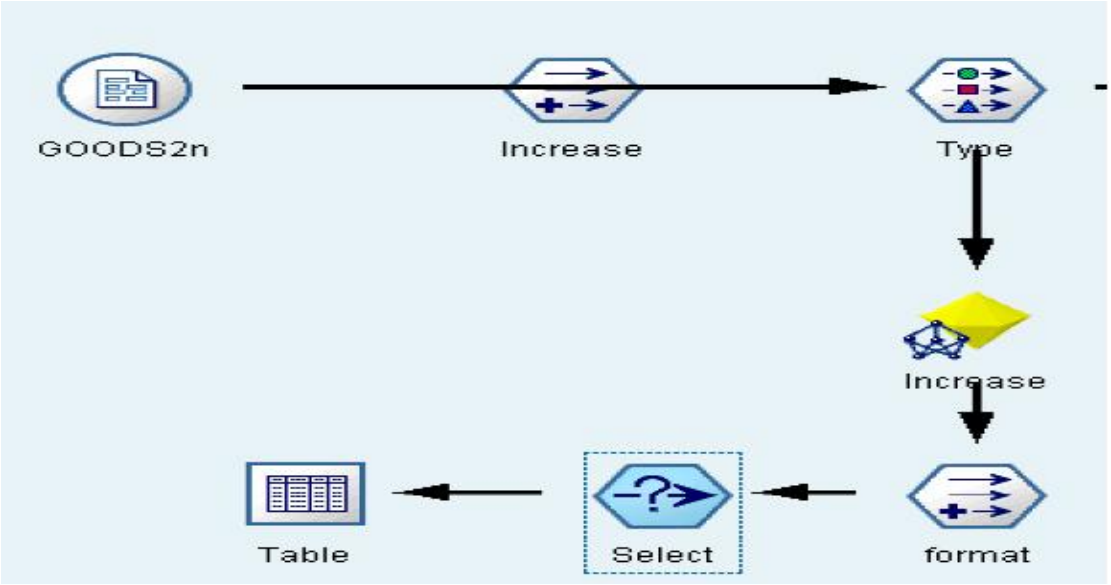
2．输出规范化

$N-Increase 栏表示促销后销售额可能增减的比率。由于神经网络的最终输出需要规范到[0,1]区间，所以我们选择输出值在(0,1)内连续的 S 形函数将结果规范化。S 型函数表达式为 $f(x)=\frac{1}{1+e^{-x}}$。我们通过增加 Derive 结点将结果其规范化。

Class	Cost	Promotion	Before	After	$N-Increase	format
Luxury	31.2...	1467	2233...	238333	5.359	0.995
Drink	82.5...	1316	1989...	219791	9.903	1.000
Luxury	10.4...	1734	2480...	266357	6.531	0.999
Drink	40.4...	1002	2159...	235013	7.865	1.000
Drink	20.2...	1127	2890...	305659	8.335	1.000
Meat	59.3...	1884	2347...	241302	4.338	0.987
Meat	71.1...	1655	2087...	216708	3.567	0.973
Drink	62.7...	1108	1922...	204458	8.668	1.000
Drink	98.2...	1075	2342...	248692	8.574	1.000
Drink	34.6...	1644	1109...	121988	11.419	1.000
Luxury	87.4...	1105	1361...	140323	4.465	0.989
Drink	92.7...	1828	2091...	239858	12.096	1.000

3．选择促销方案

根据神经网络模型的预测输出，我们可以选出 GOODS2n 文件中包含的可执行促销方案。假定预测结果经规范化后结值 1 的方案为可执行方案，我们需要增加一个结点来选出满足这些条件的结点。Clementine 为我们提供了 Select 结点，它可以从数据集中筛选出满足预定条件的记录。



从 Record OPs 栏内选择 Select 结点连接到 Format 结点后，在它的属性设置中选择包含 format＝1.000 的结点，整个流程图由下所示：

	Class	Cost	Promotion	Before	After	$N-Increase	format
1	Drink	92.760	1828	2091...	239858	12.096	1.000
2	Drink	98.150	1706	2234...	247907	11.666	1.000
3	Drink	44.540	1938	1718...	196179	12.394	1.000
4	Drink	103.3...	1904	2447...	281376	12.309	1.000
5	Drink	76.190	1888	2579...	288452	12.225	1.000
6	Drink	102.4...	1718	1537...	170421	11.782	1.000
7	Drink	53.880	1902	2354...	265308	12.254	1.000
8	Drink	40.870	1720	1434...	161531	11.697	1.000
9	Drink	48.270	1690	2009...	224928	11.540	1.000
10	Drink	87.110	1824	1233...	140228	12.148	1.000
11	Drink	26.970	1711	2261...	250797	11.536	1.000
12	Drink	36.960	1945	2545...	287404	12.335	1.000
13	Drink	70.150	1714	2001...	227041	11.679	1.000
14	Drink	90.530	1697	1252...	139851	11.720	1.000
15	Drink	50.300	1783	1555...	175344	11.932	1.000

如果我们只需要得到这些方案的某些字段，而不想知道它的全部细节，则可以在 Select 和 Table 键中增设 Filter 结点，将不需要的字段过滤。

P.S.：在神经网络示例的学习中，我们用到了Neural Net、Select结点。

六、实验报告要求

1、能够将软件操作步骤在报告中凸现出来；

2、结果的分析要合理、准确。

七、思考题

1、说明人工神经网络设计和“传统”信息处理系统设计之间的基本区别；

2、神经网络的基本组成是什么；

3、讨论前馈和循环神经网络之间的区别；

4、了解相应的 SAS 数据挖掘软件的基本操作流程。

八、注意事项

1、神经网络的结果的合理的解释；

2、软件操作应当正确且熟练掌握。