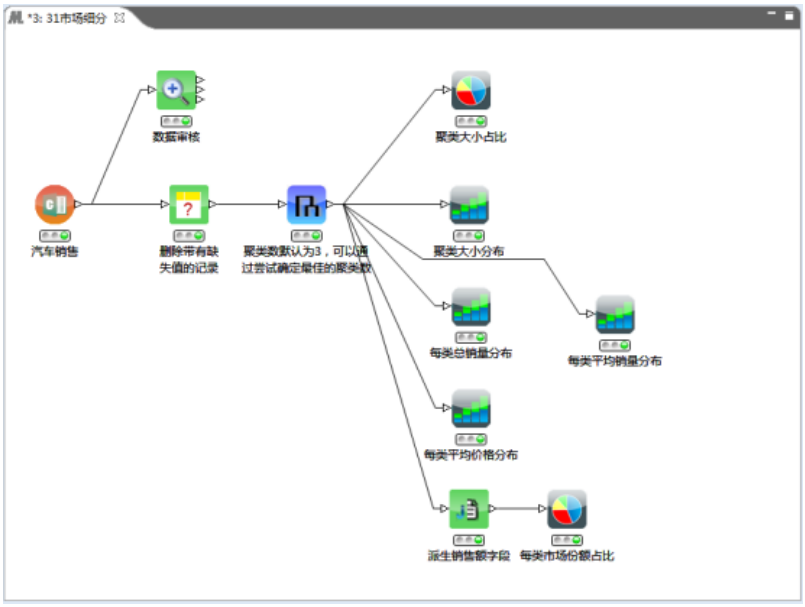
案例 6：聚类分析在企业市场细分中的运用

聚类与分类的不同在于，聚类所要求划分的类是未知的。聚类是将数据分类到不同的类或者簇这样的一个过程，所以同一个簇中的对象有很大的相似性，而不同簇间的对象有很大的相异性。从统计学的观点看，聚类分析是通过数据建模简化数据的一种方法。传统的统计聚类分析方法包括系统聚类法、分解法、加入法、动态聚类法、有序样品聚类、有重叠聚类和模糊聚类等。从机器学习的角度讲，簇相当于隐藏模式。聚类是搜索簇的无监督学习过程。与分类不同，无监督学习不依赖预先定义的类或带类标记的训练实例，需要由聚类学习算法自动确定标记，而分类学习的实例或数据对象有类别标记。聚类是观察式学习，而不是示例式的学习。聚类分析是一种探索性的分析，在分类的过程中，人们不必事先给出一个分类的标准，聚类分析能够从样本数据出发，自动进行分类。聚类分析所使用方法的不同，常常会得到不同的结论。不同研究者对于同一组数据进行聚类分析，所得到的聚类数未必一致。

从实际应用的角度看，聚类分析是数据挖掘的主要任务之一。而且聚类能够作为一个独立的工具获得数据的分布状况，观察每一簇数据的特征，集中对特定的聚簇集合作进一步地分析。聚类分析还可以作为其他算法（如分类和定性归纳算法）的预处理步骤。

聚类分析的核心思想就是物以类聚，人以群分。在市场细分领域，消费同一种类的商品或服务时，不同的客户有不同的消费特点，通过研究这些特点，企业可以制定出不同的营销组合，从而获取最大的消费者剩余，这就是客户细分的主要目的。在销售片区划分中，只有合理地将企业所拥有的子市场归成几个大的片区，才能有效地制定符合片区特点的市场营销战略和策略。金融领域，对基金或者股票进行分类，以选择分类投资风险。

下面以一个汽车销售的案例来介绍聚类分析在市场细分中的应用。



业务理解：数据名称《汽车销售.csv》。该案例所用的数据是一份关于汽车的数据，该数据文件包含销售值、订价以及各种品牌和型号的车辆的物理规格。订价和物理规格可以从 edmunds.com 和制造商处获得。定价为美国本土售价。如下：

文件表 - 3:1 - CSV(汽车销售)

文件

输出表 - 行数: 157　规格 - 列数: 14

行ID	S 名称	S 制造商	S 型号	D 销售量	D 类型	D 价格	D 引擎	D 马力	D 轴
Row0	Acura_In	Acura	Integra	282843.0	0.0	21.5	1.8	140.0	101.2
Row1	Acura_TL	Acura	TL	367335.0	0.0	28.4	3.2	225.0	108.1
Row2	Acura_CL	Acura	CL	264716.0	0.0	?	3.2	225.0	106.9
Row3	Acura_RL	Acura	RL	215036.0	0.0	42.0	3.5	210.0	114.6
Row4	Audi_A4	Audi	A4	301538.0	0.0	23.99	1.8	150.0	102.6
Row5	Audi_A6	Audi	A6	293279.0	0.0	33.95	2.8	200.0	108.7
Row6	Audi_A8	Audi	A8	32208.0	0.0	62.0	4.2	310.0	113.0
Row7	BMW_323i	BMW	323i	298300.0	0.0	26.99	2.5	170.0	107.3
Row8	BMW_328i	BMW	328i	222256.0	0.0	33.4	2.8	193.0	107.3
Row9	BMW_528i	BMW	528i	286374.0	0.0	38.9	2.8	193.0	111.4
Row10	Buick_Ce	Buick	Century	451700.0	0.0	21.975	3.1	175.0	109.0
Row11	Buick_Re	Buick	Regal	367249.0	0.0	25.3	3.8	240.0	109.0
Row12	Buick_Pa	Buick	Park Avenue	332686.0	0.0	31.965	3.8	205.0	113.8
Row13	Buick_Le	Buick	LeSabre	442193.0	0.0	27.885	3.8	205.0	112.2

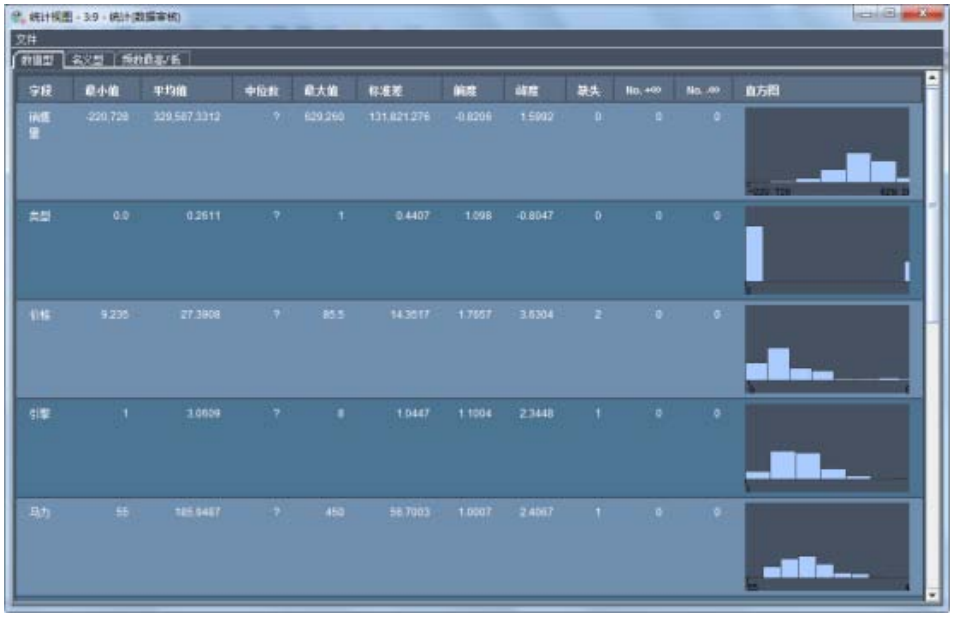
表 1：数据视图

业务目标：对市场进行准确定位，为汽车的设计和市场份额预测提供参考；

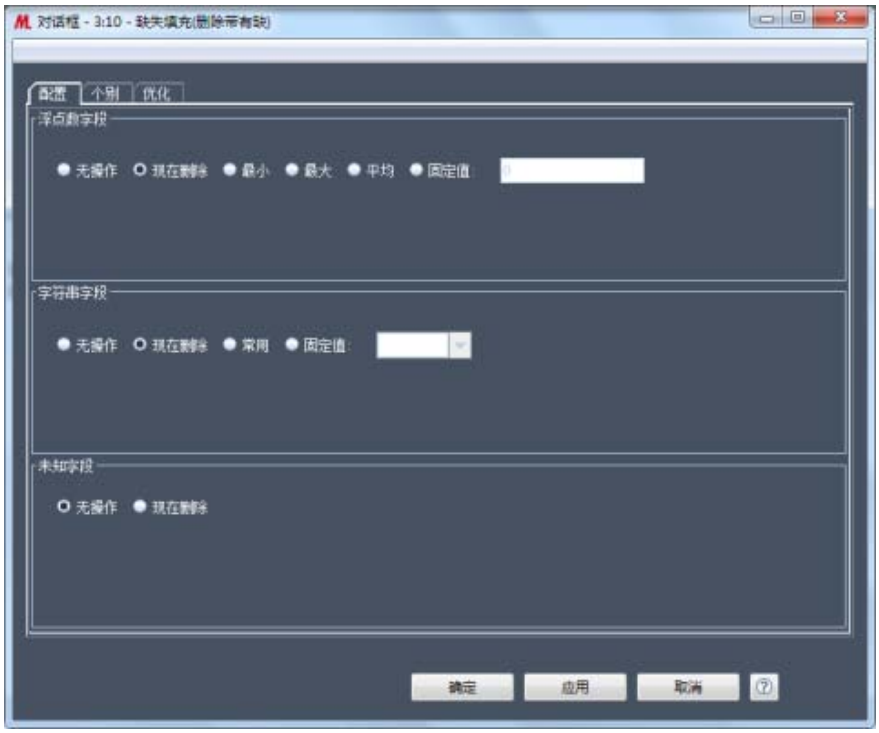
数据挖掘目标：通过聚类的方式对现有的车型进行分类。

数据准备：通过数据探索对数据的质量和字段的分布进行了解，并排除有问题的行或者列优化数据质量。

第一步，我们使用统计节点审核数据的质量，从审核结果中我们发现存在缺失的数据，如下图所示：



第二步，对缺失的数据进行处理，我们选择使用缺失填充节点删除这些记录。配置如下：



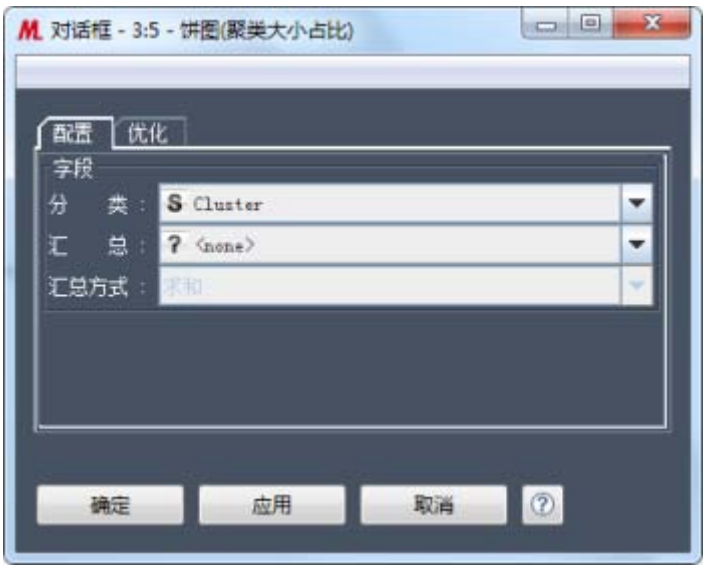
建模：我们选择层次聚类进行分析，尝试根据各种汽车的销售量、价格、引擎、马力、轴距、车宽、车长、制动、排量、油耗等指标对其分类。

因为层次聚类不能自动确定分类数量，因此需要我们以自定义的方式规定最后聚类的类别数。层次聚类节点配置如下（默认配置）：

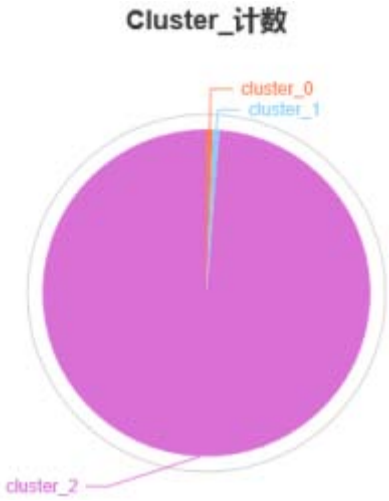
可以使用交互表或者右击层次聚类节点查看聚类的结果，如下图所示：

行ID	马力	轴距	车宽	车长	制动	排里	油耗	Cluster
Row53	300.0	110.2	70.9	189.2	3.693	19.8	21.0	cluster_0
Row69	302.0	99.0	71.3	177.1	4.125	21.1	20.0	cluster_0
Row111	160.0	100.5	67.6	176.6	3.042	15.8	25.0	cluster_0
Row114	236.0	104.9	71.5	185.7	3.601	18.5	23.0	cluster_0
Row16	275.0	108.0	75.5	200.6	3.843	19.0	22.0	cluster_0
Row54	290.0	112.2	72.0	196.7	3.89	22.5	22.0	cluster_0
Row34	163.0	103.7	69.1	190.2	2.879	15.9	24.0	cluster_0
Row82	132.0	108.0	71.0	186.3	2.942	16.0	27.0	cluster_0
Row96	124.0	102.4	66.4	176.9	2.452	12.1	31.0	cluster_0
Row60	210.0	107.1	70.3	194.1	3.443	19.0	22.0	cluster_0
Row108	115.0	98.9	68.3	163.3	2.762	14.6	26.0	cluster_0
Row27	163.0	103.7	69.7	190.9	2.967	15.9	24.0	cluster_0
Row70	190.0	94.5	67.5	157.9	3.055	15.9	26.0	cluster_0
Row107	115.0	97.4	66.7	160.4	3.079	13.7	26.0	cluster_0

再使用饼图查看每个类的大小。饼图配置如下：

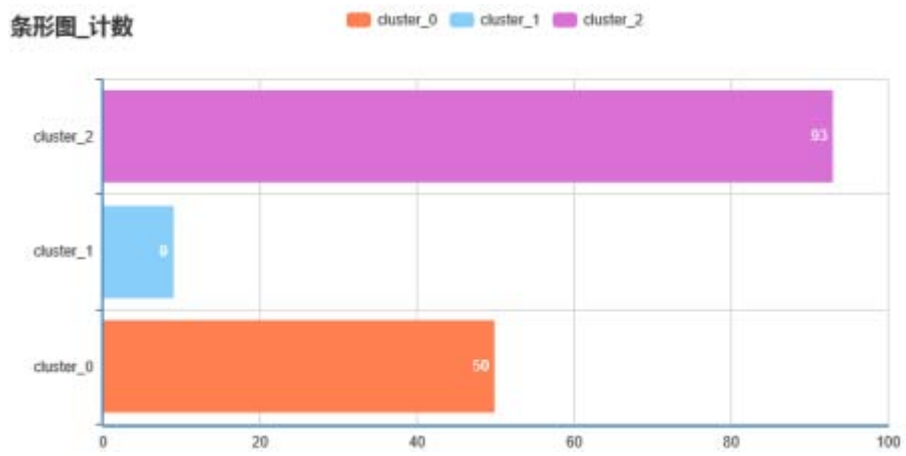


结果如下：

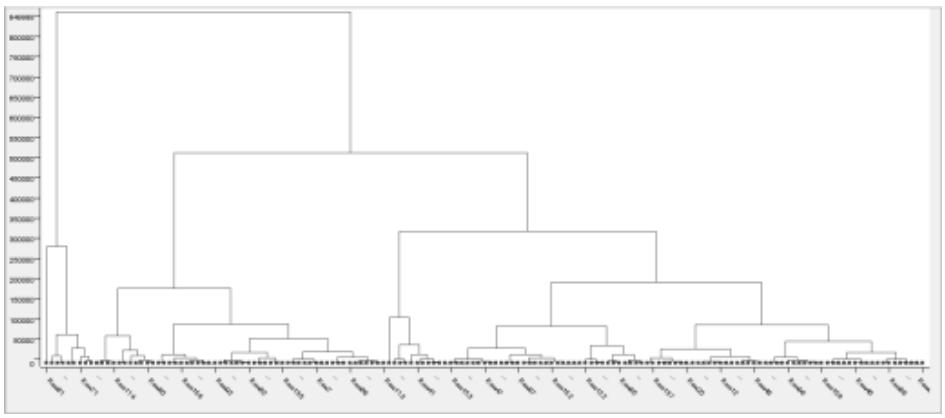


从图中可见，分成的三个类样本数差异太大，cluster_0 和 cluster_1 包含的样本数都只有 1，这样的分类是没有意义的，因此需要重新分类。我们尝试在层次聚类节点的配置中指定新的聚类方法：完全。新的聚类样本数分布如下：

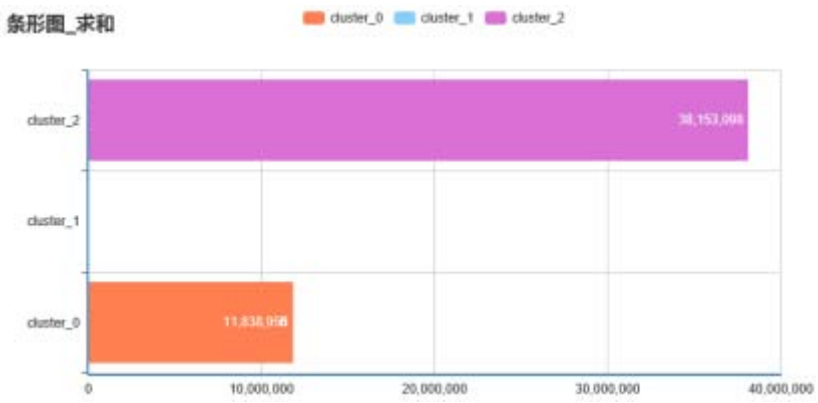
cluster_0、 cluster_1、cluster_2 的样本数分别为：50、9、93。



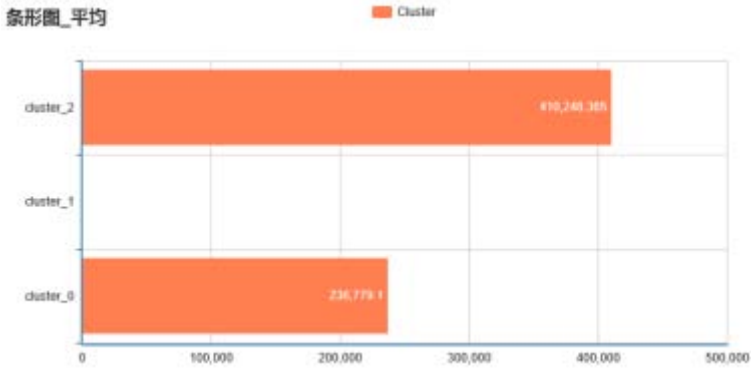
执行后输出树状/冰柱图，可以从上往下看，一开始是一大类，往下走就分成了两类，越往下分的类越多，最后细分到每一个记录是一类，如下所示：



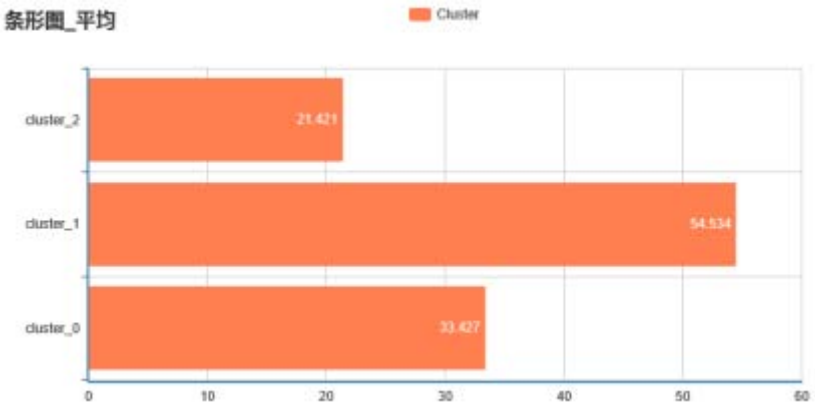
我们可以再使用条形图查看每类的销售量、平均价格，如下图所示：



每类总销量分布图



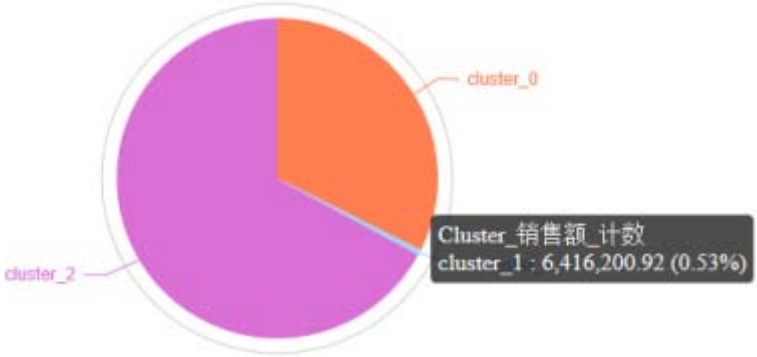
每类平均销量分布图



每类平均价格分布图

我们再看一下每类的销售额分布情况。首先，我们需要使用 Java 代码段节点或者派生节点生成销售额字段，配置如下：

再使用饼图查看销售额分布情况，cluster_0、 cluster_1、cluster_2 的市场份额分别为：32.39%、0.53%和 67.08%，如下图所示：



小结

通过这个案例，大家可以发现聚类分析确实很简单。进行聚类计算后，主要通过图形化探索的方式评估聚类合理性，以及在确定聚类后，分析每类的特征。

注：案例来源于http://wiki.smartbi.com.cn

思考题：（1）案例中使用了那些聚类分析方法？效果如何？

（2）通过此案例，你能总结出聚类分析的基本步骤吗？