## （一）客户分层标签体系

基于基础数据构建客户指标体系，从体系的各个维度构建聚类模型，结合业务目标和规则，对聚类结果进行解读，形成新的综合标签。

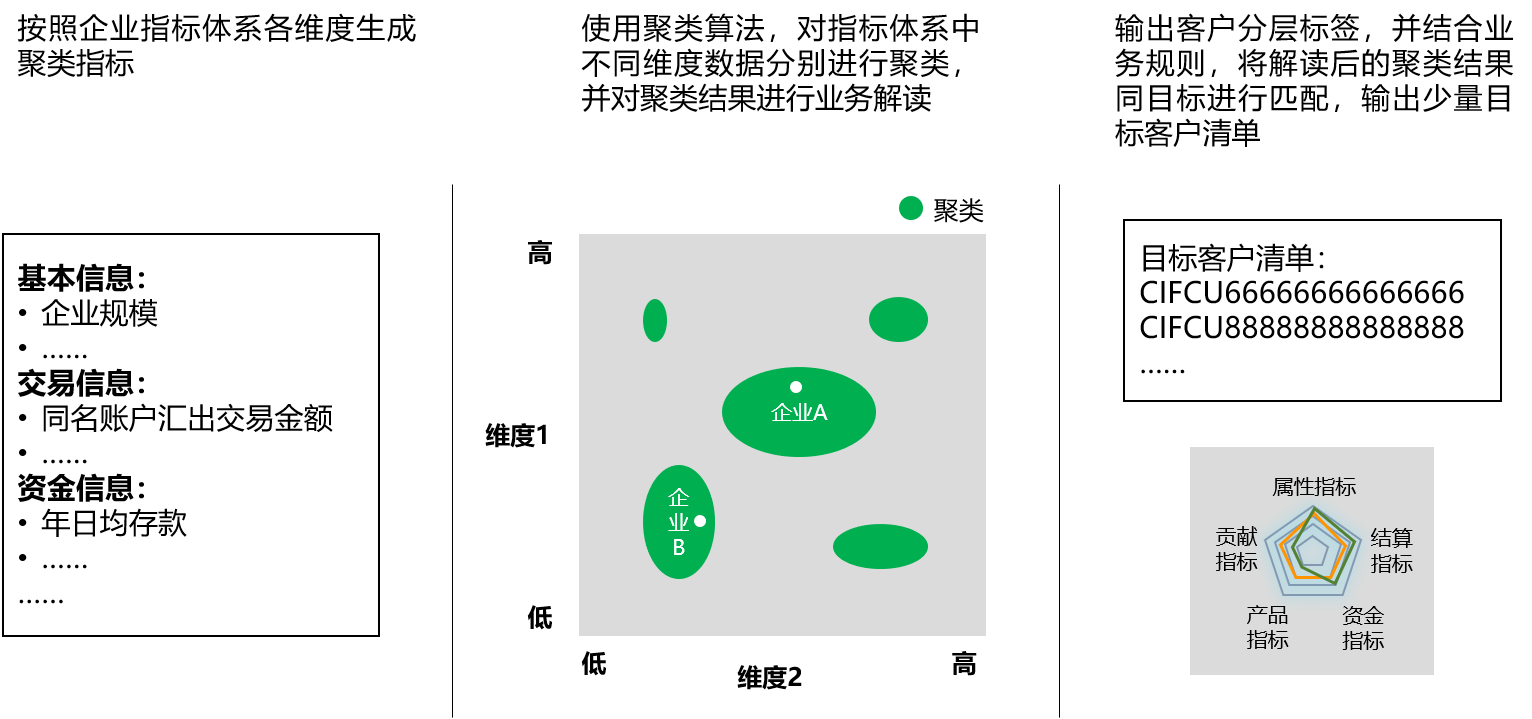


图1 分层标签体系

## （二）聚类分析方法论

（1）数据处理

* 缺失值检验：删除缺失率大于95%的指标，对于缺失率大于80%的指标，生成指示变量。
* 常数项检验：检验出数值为常数的特征并删除，因为此类特征在企业间没有差异性，对聚类模型无用。
* 共线性检验：找出与其他特征线性相关的特征，将其删除，降低特征数量，提高聚类效率。
* 去除异常值：将大于99百分位的特征值转化为99百分位数，将小于1百分位的特征值转化为1百分位数，从而降低异常值对聚类模型的影响。
* 缺失值补0处理。

（2）特征降维

当变量过多时，聚类模型训练速度缓慢，并且因变量过多，模型难以解释，所以需要进行特征选取或特征聚类，达到降维的目的。

方法一：

采用PCA等降维方法，将特征映射到更低纬度的特征，然而新生成的变量极难归纳出业务意义，不利于最终聚类结果的解读。

方法二：

变量聚类，通过聚类时的重要性，挑选出每一类最具有代表性的变量，因考虑到此类方法选出的特征业务含义容易解读，便于结合业务规则进行目标客户挖掘，故拟选用本方法。

（3）模型选择

决定入模变量后，通过调节参数，将数据送入不同的模型，通过手肘法和轮廓系数法确定聚类模型的最佳分类个数，对比最后分类结果选出分类占比较均匀且有区分度的模型作为最后的模型。

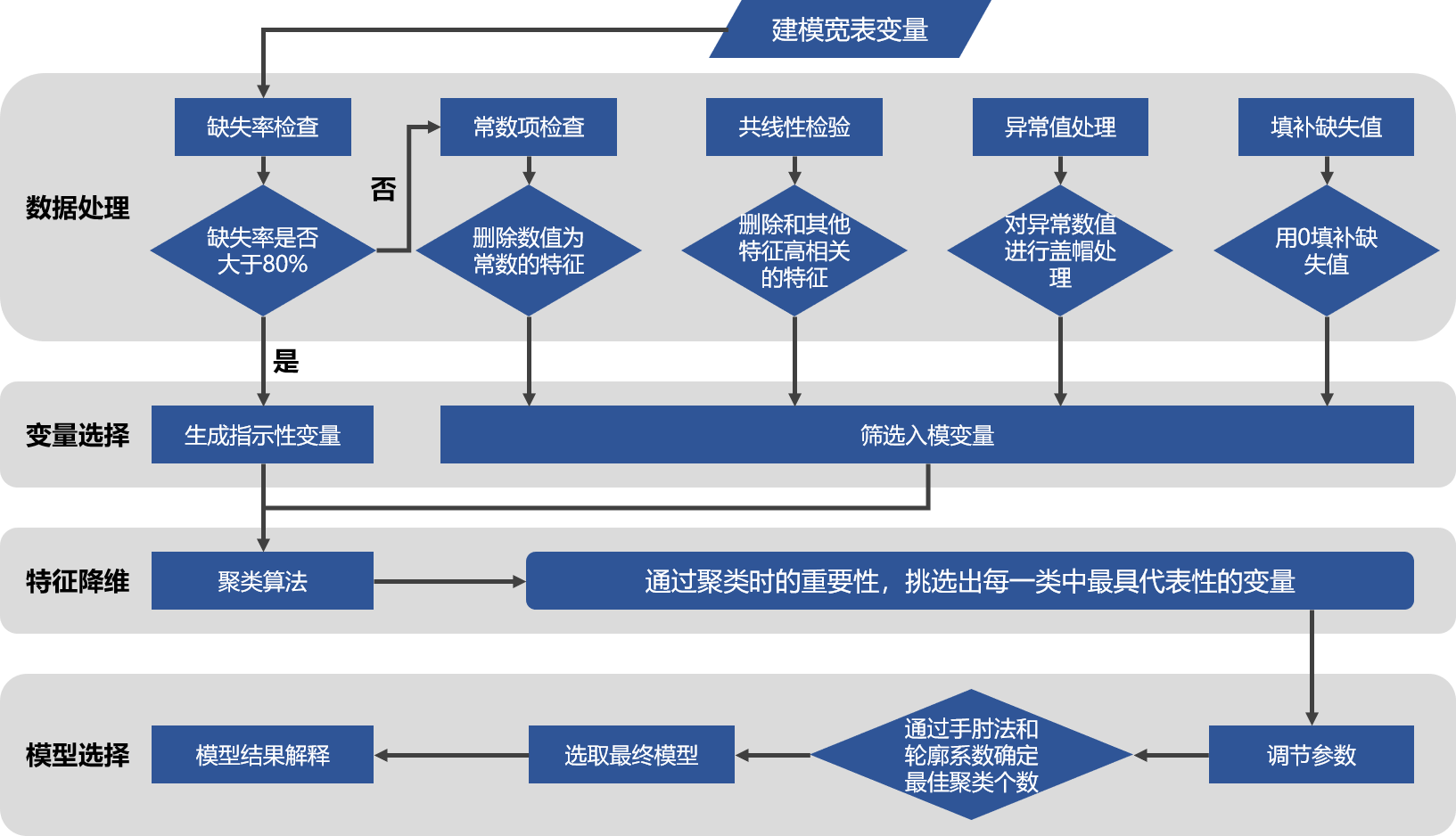


图2 聚类分析方法论