**PMML聚类模型导入导出**

目录

[一、 PMML聚类模型介绍 2](#_Toc434914720)

[二、 聚类模型结构 2](#_Toc434914721)

[三、 构造导出PMML过程 3](#_Toc434914722)

[四、 解析导入PMML过程 3](#_Toc434914723)

[五、 其他聚类算法 3](#_Toc434914724)

# PMML聚类模型介绍

* 1. PMML模型中肯定包括数据字典，字典中包括数据输入列、类型以及可能取值范围。
  2. 在PMML中添加聚类模型信息，有提供者提供PMML信息，获取者解析PMML中包含的信息，并创建聚类模型，完成聚类工作。
  3. 通过PMML模型，可以做到跨平台，跨软件的挖掘模型信息传递。

# 聚类模型结构

* 1. ClusteringModel是用于标记聚类模型的标签。
  2. ClusteringModel中包括modelName、functionName、algorithmName、modelClass、聚类个数、是否可产生数值等属性。modelName表示模型名，是可选的；functionName表示挖掘方法，这个属性是必须的，在聚类模型中必须被设置为clustering；algorithmName表示算法名称，可选；modelClass是必选的，表示模型类别，分为基于中心的和基于分布的两种，基于中心模型如K-Means，基于分布的模型如EM。
  3. MiningSchema
     1. 按照PMML文档的规范，MiningSchema是模型元素的门卫，哪个模型都要有，其中包含的MiningField应该取至DataDictionary的DataField，用于标识本模型真正使用到的属性。
  4. Output
     1. 定义输入格式，OutputField定义每列输入的信息。
  5. LocalTransformations
     1. 类似于TransformationDictionary。
  6. ComparisonMeasure
     1. 指定度量指标，如euclidean、squaredEuclidean、minkowski等，是必选的。
  7. ClusteringField
     1. 指定参与聚类的列。
  8. Cluster
     1. 每个Cluster包含一个聚类的簇信息

# 构造导出PMML过程

* 1. 这里使用的是org.dmg.pmml包，首先创建PMMLDocument，通过PMMLDocument.getPMML()获取PMML，使用PMML.addNew方法按照dmg.org网站上提供的schema.xsd信息进行添加，最终完成PMMLDocument的构造。

# 解析导入PMML过程

* 1. 在导入PMML后，首先需要经过PMMLDocument.validate()的验证，如果词法和语法有问题，就不能保证解析出的内容完整性。
  2. 获取PMMLDocument对象后，通过ClusteringField获取参与聚类的列，获取指标和簇心信息。
  3. 比如K-Means，会获取各个质点的信息，然后根据ComparisonMeasure指定的指标类型，计算测试集各点所属聚类信息。

# 其他聚类算法

* 1. 对于一些聚类算法，PMML并没有直接的支持，比如modelClass中不支持密度聚类。
  2. 这时，可以使用ClusteringModel.addNewExtension()添加扩展，使用Extension.newCursor获取XmlCursor，通过XmlCursor添加扩展内容。在导入模型时，再通过XmlCursor解析扩展内容，来达到定制的目的。