学习 › Open source

何为 PMML?

探索预测分析和开放式标准的强大功能

Alex Guazzelli
2013 年 1 月 25 日发布

PMML 简介

如果现在有人问您是否使用过预测分析，您可能会回答“没有”。其实并非如此，您可能
知。当您刷信用卡或在网上使用信用卡时，一个预测分析模型检查这笔交易是否是欺诈
是一个预测分析模型为您推荐了一部特别的电影。事实上预测分析已成为我们生活的一
的帮助。

随着桥梁、建筑、工业生产流程和机械传感数据的生成，预测解决方案必定可以提供一
故障和问题发生之前对您提出警告。传感器还可用于监控人类，如应用于特护病房。IBM
Technology 现在正在合作实现一个用于监控早产儿的数据分析和预测解决方案，其中采
测危急生命的传染。

但是仅预测分析可以起作用吗？视情况而定。开放式标准是其中最重要的组成部分。要
析带来的益处，系统和应用程序需要通过下列标准轻松交换信息。PMML 支持在应用程

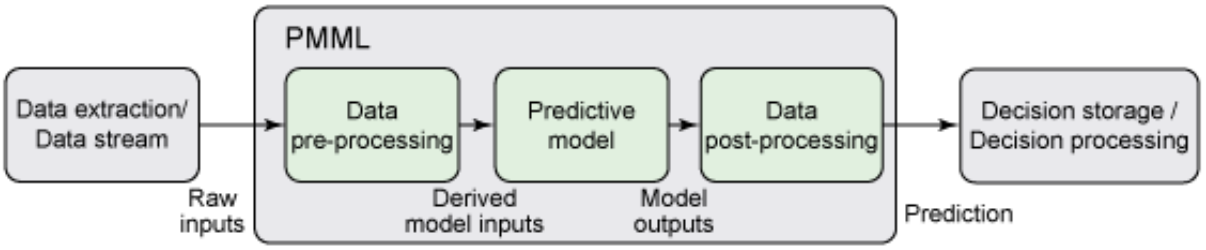
主要分析供应商对 PMML 的采用是支持互操作性公司中的典型例子。IBM、SAS、Micro
都是数据挖掘群组（Data Mining Group，DMG）中的成员，DMG 是使 PMML 成形的委
公司同样是该委员会的成员。PMML 可以塑造预测分析世界，并使其成为一个对您来说

PMML 基础知识

PMML 是一种事实标准语言，用于呈现数据挖掘模型。预测分析模型 和数据挖掘模型 是技术了解大量历史数据中隐藏的模式。预测分析模型采用定型过程中获取的知识来预测新在不同的应用程序之间轻松共享预测分析模型。因此，您可以在一个系统中定型一个模型移动到另一个系统中，并在该系统中使用上述模型预测机器失效的可能性等。

PMML 是数据挖掘群组的产物，该群组是一个由供应商领导的委员会，由各种商业和开放接）组成。因此，现在的大部分领先数据挖掘工具都可以导出或导入 PMML。作为一个已以呈现用于从数据中了解模型的统计技术（如人工神经网络和决策树），也可以呈现原始理（参见 图 1）。

图 1. PMML 包含数据预处理和数据后处理以及预测模型本身



PMML 文件的结构遵从了用于构建预测解决方案的常用步骤，包括：

1. 数据词典，这是一种数据分析阶段的产品，可以识别和定义哪些输入数据字段对于解数值、顺序和分类字段。
2. 挖掘架构，定义了处理缺少值和离群值的策略。这非常有用，因为通常情况，当将模可能为空或者被误呈现。
3. 数据转换，定义了将原始输入数据预处理至派生字段所需的计算。派生字段（有时也或修改，以获取更多相关信息。例如，为了预测停车所需的制动压力，一个预测模型雨？）作为原始数据。派生字段可能会将这两个字段结合起来，以探测路上是否结冰来预测停车所需的制动压力。
4. 模型定义，定义了用于构建模型的结构和参数。PMML 涵盖了多种统计技术。例如，的神经层和神经元之间的连接权重。对于一个决策树来说，它定义了所有树节点及简
5. 输出，定义了预期模型输出。对于一个分类任务来说，输出可以包括预测类及与所有
6. 目标，定义了应用于模型输出的后处理步骤。对于一个回归任务来说，此步骤支持将数（预测结果）。
7. 模型解释，定义了将测试数据传递至模型时获得的性能度量标准（与训练数据相对）

矩阵、增益图及接收者操作特征（ROC）曲线图。

8. 模型验证，定义了一个包含输入数据记录和预期模型输出的示例集。这是非常重要的
型时，该模型需要通过匹配测试。这样就可以确保，在呈现相同的输入时，新系统可
情况是这样的话，一个模型将被认为经过了验证，且随时可用于实践。

考虑到 PMML 支持预测解决方案被整体表达（包括数据预处理、数据后处理和建模技术
骤的反映。

互操作性：在应用程序之间共享解决方案

在应用程序之间共享模型是预测分析成功的关键。但是，要共享模型，首先必须建立一个

建模

建模由几个阶段组成，其中包括一个彻底数据分析阶段。在此阶段，您可以对原始数据
最重要信息（这将产生上述 步骤 1 中所定义的数据词典）。您也可以创建派生字段，采
（步骤 3）。然后原始和派生字段即可用于模型定型。这个过程的结果就是，您在分析阶
构建最终模型（步骤 4）。构建模型后，将根据测试数据集测试模型性能（步骤 7）。整
解决的问题的复杂程度而定。通常情况下，您可以建立多个模型，有时可使用不同的统计
最终模型则可能包含一种单一技术或者多种技术的组合并产生一个包含多个模型的 PMM

模型部署

模型部署是将预测解决方案有效地应用于实践的过程，这项任务一般由与建模过程相分离
解决方案要监控的系统和流程紧密集成。但是由于可用的快速 Internet 连接，这些系统无
Internet 的 web 服务轻松实现整合。在这种情况下，您可以从云计算中获益，根据需要
量数据。

将一个预测分析模型应用于实践中时，您会期望它可以持续几个月甚至几年完成自己的任
更新。在这种情况下，需要建立和部署另外一个模型，代替原来的模型。但是，通常情况
须关注互操作性和开放式标准需求。

模型共享

如果没有 PMML 这样的语言，部署预测解决方案可能会非常困难繁琐，因为不同的系统
个系统移至另一个系统时，您必须经历一个漫长的、容易出现错误和误呈现的翻译过程。
了。最近，我非常惊讶地发现，一个大型金融企业花了半年至一年的时间部署其数据挖
PMML， 您只需几分钟就可以完成部署。

PMML 支持从应用程序 A 到 B 再到 C 轻松共享预测解决方案，并在建模阶段结束后立即
IBM SPSS Statistics 中建立了一个模型，然后在 ADAPA（Zementis 预测决策平台）中进
见 参考资料 中的链接）。或者您也可以将该模型移动至 IBM InfoSphere™，其中模型将
将模型移动到 KNIME，这是一个构建和可视化来自德国 University of Konstanz 的数据流
功能：支持在应用程序之间真正实现模型和解决方案的互操作性。PMML 还支持您向最
性。现在，您可以直接在 Microsoft® Office Excel 中从之前部署到 Zementis ADAPA 平台
然后单击 **Score**。

接下来，我将以一个名为 *predictive maintenance* 的字段为例，举例说明预测分析和 PM

预见性维护：PMML 和数据挖掘的应用

预见性维护，顾名思义就是在故障和事故发生之前对材料或流程进行维护或改变，这是一
报告桥梁和建筑等结构以及能量转换器、水泵、气泵、闸和阀等机械现状的小型经济高效
现。

我非常高兴曾参与一个涉及提前检测转动设备故障的项目。如果没有预见性维护，您必
生产线中，这意味着必须停止整个生产线运行，直到机器被修好或更换为止。有了预见性
出故障的设备，如在小量生产期间或者计划维护周期内进行维修或替换。为了提前检测
战。原始输入数据仅包含每小时内几秒钟捕获的振动信号。考虑到很多转动设备（和传感
号质量由于邻近设备的干扰会有所降低。

尽管存在干扰问题，我们仍然可以使用数据挖掘和分析成功抵消噪音。为此，我们主要
具包。然后我们使用 IBM SPSS Statistics 建立了多个模型。最终模型是一个神经网络，
决方案在 PMML 中完全呈现，我们轻松地在 Zementis ADAPA 平台（我们之前已在客户
我们就集中精力解决剩下的挑战：确保传感器输入可以如期传入我们的解决方案。我们
制造现场上实施的维护过程和指导方针的一部分。

使用预测分析模型作为监督工具，您可以预防意外事故的发生。通过在事故发生之前对
全的环境。对于石油化工业来说，预测分析可以而且必须作为石油钻探和勘查安全措施

导出 PMML

可以通过多种统计工具轻松导出 PMML。如上所述，顶级分析公司导出并导入其产品 PM
Statisticsis 中，您可以在选择好所有合适模型参数后选择将模型导出为 XML 文件（PMM
型。对于一个神经网络模型来说，典型参数对网络中使用的层和神经元的数量进行了说明
前，选择 **Export** 选项卡保存模型。将您的解决方案保存为 PMML 文件是不错的实践，即
成最终模型之前保存所有试验模型的 PMML 记录。您和团队中的其他人员都可以使用这

PMML 深度解析

您已了解了何为 PMML 及其重要性，现在让我们来深入探究这种语言本身。如上所述，P
方案的八大步骤，从在 “数据词典” 步骤中定义原始输入数据字段到在 “模型验证” 步骤中

清单 1 展示了一个含有三个字段的解决方案中 PMML 元素 `DataDictionary` 的定义，这
类输入字段 `Element` 和数值型输出字段 `Risk`。

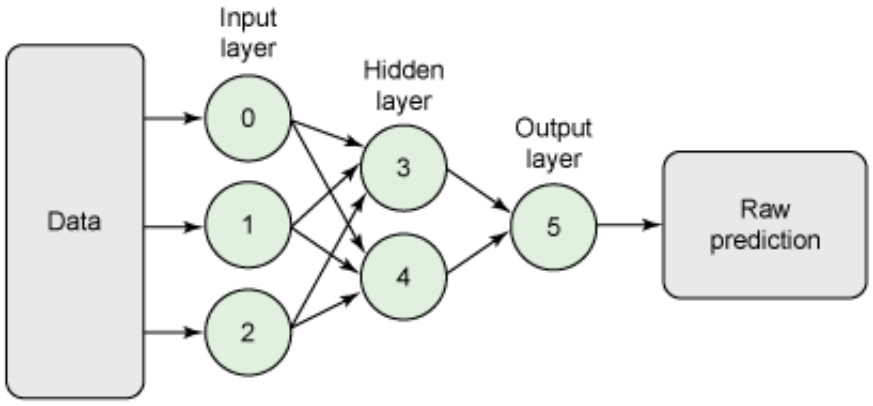
清单 1. `DataDictionary` 元素

```
<DataDictionary numberOfFields="3">
    <DataField dataType="double" name="Value" optype="continuous">
        <Interval closure="openClosed" rightMargin="60" />
    </DataField>
    <DataField dataType="string" name="Element" optype="categorical">
        <Value property="valid" value="Magnesium" />
        <Value property="valid" value="Sodium" />
        <Value property="valid" value="Calcium" />
        <Value property="valid" value="Radium" />
    </DataField>
    <DataField dataType="double" name="Risk" optype="continuous" />
</DataDictionary>
```

请注意，对于字段 `Value`，范围从负无穷大到 60 的值是有效值。高于 60 的值被定义为
用 PMML 元素 `MiningSchema` 为无效值和遗漏值定义合适的处理方法。）考虑到字段 `El`
如果该特定字段的数据提要包含元素 `Iron`，将该元素作为无效值处理。

图 2 展示了神经网络模型的图形表示，其中输入层包含 3 个神经元，隐藏层包含 2 个神
望的，PMML 可以完全呈现这样一个结构。

图 2. 一个简单的神经网络模型，其中在对预测进行计算之前，数据经过一系列层



清单 2 展示了隐藏层及其神经元以及输入层（0、1 和 2）和隐藏层（3 和 4）中神经元的

清单 2. 在 PMML 中定义神经层及其神经元

```
<NeuralLayer numberOfNeurons="2">
    <Neuron id="3" bias="-3.1808306946637">
        <Con from="0" weight="0.119477686963504" />
        <Con from="1" weight="-1.97301278112877" />
        <Con from="2" weight="3.04381251760906" />
    </Neuron>
    <Neuron id="4" bias="0.743161353729323">
        <Con from="0" weight="-0.49411146396721" />
        <Con from="1" weight="2.18588757615864" />
        <Con from="2" weight="-2.01213331163562" />
    </Neuron>
</NeuralLayer>
```

PMML 不是一件艰难的事。其复杂程度反映了其呈现的建模技术的复杂程度。事实上，它
密和黑匣子。利用 PMML，任何预测解决方案都可以采用同样的顺序用同一种语言元素呈

在公司中，PMML 不仅可以作为应用程序之间也可以作为部门、服务提供商及外部供应商
就成为定义预测解决方案交流的单一、清晰流程的一个标准。

结束语

PMML 支持预测解决方案的即时部署。它是呈现预测分析模型的事实标准，目前受所有顶
的传感器以及生成数据的增多，PMML 等预测分析和开放式标准是使一切有意义的关键。
案和预见性维护仅仅是其中的部分例子。所以，挽起袖口开始投入 PMML 的工作中去吧

相关主题

- PMML in Action: Unleashing the Power of Open Standards for Data Mining and Prec Ching Lin、Tridivesh Jena；CreateSpace，2010 年 5 月）：从实践角度探讨 PMML
- 数据挖掘群组（DMG）：从该独立、供应商引领、开发预测模型标记语言（PMML）源。
- Zementis PMML Resources 页面：查看完整的 PMML 示例，包括集群模型、决策树、归模型、记分卡和支持向量机。
- Wikipedia 中的 PMML 页面：了解 PMML 及规范等信息链接的概览。
- Wikipedia 中的 预测分析页面：阅读统计分析领域中常见的类型、应用程序和统计技
- Wikipedia 中的 数据挖掘页面：访问并阅读从数据中提取模式的流程的更多信息。
- IBM SPSS Statistics 18（即之前的 SPSS Statistics）：掌握高级统计分析的强大功能 是一名初学者，它所提供的一整套工具都可以满足您的需求。
- ADAPA：试用革命性预测分析决策管理平台，可将其作为云计算服务或现场服务提供 的环境，用于部署您的数据挖掘模型和业务逻辑，并将它们应用于实践。
- IBM WebSphere Application Server：在利用 IBM WebSphere Application Server 降 建、部署和管理所有健壮、灵活和可重用的 SOA 业务应用程序和服务。
- IBM 产品评估版本：下载或 在线试用 IBM SOA Sandbox，并开始使用来自 DB2®、L WebSphere® 的应用程序开发工具和中间件产品。
- 在 developerWorks Information Management 专区，了解关于信息管理的更多信息， 下载、产品信息以及其他资源。
- 随时关注 developerWorks 技术活动和网络广播。
- 访问 developerWorks Open source 专区获得丰富的 how-to 信息、工具和项目更新以 放源码技术进行开发，并将它们与 IBM 产品结合使用。

软件下载

Code patterns