**数据建模的经验整理**

[**1. 数据挖掘是一个过程，每一步的核心是业务知识，每一步都为了达到业务目标。 2**](#_Toc428517193)

[**2.数据挖掘中的“数据预处理“阶段比其他任何数据挖掘过程都要重要。 3**](#_Toc428517194)

[**3.不断试验是寻找正确算法、正确模型的唯一方案。 5**](#_Toc428517195)

[**4. 数据挖掘增大对业务的认知。 7**](#_Toc428517196)

[**5. 预测模型预测分析提高了信息泛化能力 8**](#_Toc428517197)

[**6.数据挖掘的结果的价值不只取决于模型的稳定性或预测的准确性。 9**](#_Toc428517198)

[**7.所有的模式（业务目标、数据挖掘目标、预测模型）因业务变化而变化。 10**](#_Toc428517199)

[**总结 11**](#_Toc428517200)

数据挖掘是利用业务知识从数据中发现和解释知识（或称为模式）的过程，这种知识是以自然或者人工形式创造的新知识。

当前的数据挖掘形式，是在20世纪90年代实践领域诞生的，是在集成数据挖掘算法平台发展的支撑下适合商业分析的一种形式。也许是因为数据挖掘源于实践而非理论，在其过程的理解上不太引人注意。20世纪90年代晚期发展的数据挖据建模标准流程，逐渐成为数据挖掘过程的一种标准化过程，被越来越多的数据挖掘实践者成功运用和遵循。

虽然数据挖据建模标准流程能够指导如何实施数据挖掘，但是它不能解释数据挖掘是什么或者为什么适合这样做。在本文中我将阐述几种准则或“定律”（其中大多数为实践者所熟知）以及另外其它一些熟知的解释。开始从理论上（不仅仅是描述上）来解释数据挖掘过程，个人观点，感兴趣可以一起研究探讨。

## 1. 数据挖掘是一个过程，每一步的核心是业务知识，每一步都为了达到业务目标。

首先要理解数据挖掘不是一种技术，是一个过程。业务目标定义了数据挖掘的主题，数据挖掘关注解决业务问题和实现业务目标，没有业务目标就不会有数据挖掘。

业务知识是数据挖掘步骤每一步的核心，有些研究者认为业务知识仅仅作用于数据挖掘过程开始阶段和最后结果实施阶段的想法是不正确的、片面的。

为了方便理解，我使用数据挖据建模标准流程阶段来说明：

1. 商业理解：商业理解必须基于业务知识，所以数据挖掘目标必须是业务目标的映射（这种映射也基于数据知识和数据挖掘知识）
2. 数据理解： 数据理解使用业务知识理解与业务问题相关的数据，以及它们是如何相关的；
3. 数据预处理：数据预处理就是利用业务知识来塑造数据，使得业务问题可以被提出和解答，通俗一点来说就是将业务知识进行再加工，类似DI Designer中的转换步骤。
4. 建模：建模是使用数据挖掘算法创建预测模型，同时解释模型和业务目标的特点，也就是说理解它们之间的业务相关性
5. 评估：评估是模型对理解业务的影响
6. 实施：实施是将数据挖掘结果作用于业务过程；

总之，没有业务知识，数据挖掘过程的每一步都是无效的，也没有“纯粹的技术”步骤。 业务知识指导过程产生有益的结果，并使得那些有益的结果得到认可。数据挖掘是一个反复的过程，业务知识是它的核心，驱动着结果的持续改善。

这背后的原因可以用“鸿沟的表现”（chasm of representation）来解释（Alan Montgomery在20世纪90年代对数据挖掘提出的一个观点）。Montgomery指出数据挖掘目标涉及到现实的业务，然而数据仅能表示现实的一 部分；数据和现实世界是有差距（或“鸿沟”）的。在数据挖掘过程中，业务知识来弥补这一差距，在数据中无论发现什么，只有使用业务知识解释才能显示其重要性，数据中的任何遗漏必须通过业务知识弥补。只有业务知识才能弥补这种缺失，这是业务知识为什么是数据挖掘过程每一步骤的核心的原因。

## 2.数据挖掘中的“数据预处理“阶段比其他任何数据挖掘过程都要重要。

数据挖掘有句著名的格言：数据预处理比数据挖掘其他任何一个过程都重要，数据预处理阶段可能占用数据挖掘过程一半甚至更多的时间，最简单的解释可以概括为“数据是困难的”，有些研究者认为采用自动化技术将会有效的减轻这个“问题”的数据获取、数据清理、数据转换等数据预处理各部分的工作量。虽然自动化技术（支持者相信这项技术可以减少数据预处理过程中的大量的工作量）是有益的，但自动化技术始终无法替代人力，无法定制指定数据。数据预处理的目的是把数据挖掘问题转化为格式化的数据，使得分析技术（如数据挖掘算法）更容易利用它。数据任何形式的变化，包括行过滤、增加减少列、行变换、列变换、值替换等（有些节点Data Studio还没有添加），这意味着问题空间的变化，因此这种分析必须是探索性的，这是数据预处理重要的原因，并且在数据挖掘过程中占有如此大的工作量，这样数据挖掘者可以从容地操纵问题空间，使得容易找到适合分析他们的方法。

有两种方法“塑造”这个问题空间。第一种方法是将数据转化为可以分析的完全格式化的数据，比如，大多数数据挖掘算法需要单一表格形式的数据，一个记录就是一个样例。数据挖掘者都知道什么样的算法需要什么样的数据形式，因此可以将数据转化为一个合适的格式。第二种方法是使得数据能够含有业务问题的更多的信息，例如，某些领域的一些数据挖掘问题，数据挖掘者可以通过业务知识和数据知识知道这些。 通过这些领域的知识，数据挖掘者通过操纵问题空间可能更容易找到一个合适的技术解决方案。

因此，通过业务知识、数据知识、数据挖掘知识从根本上使得数据预处理更加得心应手。 数据预处理的这些方面并不能通过简单的自动化实现。

需要特别指出两点：

首先是虽然经过数据获取、清理、融合等方式创建一个数据仓库，除非极特殊场景，否则数据预处理仍然是必不可少的，仍然占有数据挖掘过程一半以上的工作量。此外，即使经过了主要的数据预处理阶段，在创建一个有用的模型的反复过程中，进一步的数据预处理也是有必要的。

其次数据预处理阶段在数据挖掘过程中耗时，因为数据挖掘过程经常需要定制数据，所以数据预处理阶段在数据挖掘过程中显得尤为重要，重要不代表必须，只要能够完成指定的挖掘过程，实现业务目标，我们完全可以不去定制数据。

## 3.不断试验是寻找正确算法、正确模型的唯一方案。

机器学习有一个原则：如果我们充分了解业务知识、业务目标、初始数据、数据预处理步骤（问题空间），我们可以选择或设计一个找到最优方案的最有效的算法。一个卓越算法的参数依赖于数据挖掘问题空间一组特定的属性集，这些属性可以通过分析发现或者算法创建。但是，这种观点来自于一个错误的思想，在数据挖掘过程中数据挖掘者将问题公式化，然后利用算法找到解决方法。事实上，数据挖掘者将问题公式化和寻找解决方法是同时进行的，算法仅仅是帮助数据挖掘者的一个工具。

有五种因素说明试验对于寻找数据挖掘解决方案是必要的：

1. 数据挖掘项目的业务目标定义了兴趣范围（定义域），业务目标反映了这一点；
2. 与业务目标相关的数据及其相应的数据挖掘目标是在这个定义域上的数据挖掘过程产生的；
3. 这些过程受规则限制，而这些过程产生的数据反映了这些规则；
4. 在这些过程中，数据挖掘的目的是通过模式发现技术（数据挖掘算法）和可以解释这个算法结果的业务知识相结合的方法来揭示这个定义域上的规则；
5. 数据挖掘需要在这个域上生成相关数据，这些数据含有的模式不可避免地受到这些规则的限制。

在这里强调一下最后一点，在数据挖掘中改变业务目标，数据挖据建模标准流程不是下一个步骤仅接着上一个步骤的“瀑布”式的过程。事实上，在项目中的任何地方都可以进行任何数据挖掘过程步骤，同样商业理解也可以存在于任何一个步骤。业务目标不是简单地在开始就给定，它贯穿于整个过程。这也许可以解释一些数据挖掘者在没有清晰的业务目标的情况下开始项目，他们知道业务目标也是数据挖掘的一个结果，不是静态地给定。

Wolpert的“没有免费的午餐”理论已经应用于机器学习领域，无偏的状态好于（如一个具体的算法）任何其他可能的问题出现的平均状态。这是因为，如果我们考虑所有可能的问题，他们的解决方法是均匀分布的，以至于一个算法对一个子集是有利的，而对另一个子集是不利的。这与数据挖掘者所知的具有惊人的相似性，没有一个算法适合每一个问题。但是经过数据挖掘处理的问题或数据集绝不是随机的，也不是所有可能问题的均匀分布，他们代表的是一个有偏差的样本，那么为什么要应用NFL（没有免费的午餐）的结论？答案涉及到上 面提到的因素：问题空间初始是未知的，多重问题空间可能和每一个数据挖掘目标相关，问题空间可能被数据预处理所操纵，模型不能通过技术手段评估，业务问题本身可能会变化。由于这些原因，数据挖掘问题空间在数据挖掘过程中展开，并且在这个过程中是不断变化的，以至于在有条件的约束下，用算法模拟一个随机选择的数据集是有效的。对于数据挖掘者来说：没有免费的午餐。

这大体上描述了数据挖掘过程。但是，在有条件限制某些情况下，比如业务目标是稳定的，数据和其预处理是稳定的，一个可接受的算法或算法组合可以解决这个问题。在这些情况下， 一般的数据挖掘过程中的步骤将会减少。但是，如果这种情况稳定是持续的，数据挖掘者的午餐是免费的，或者至少相对便宜的。像这样的稳定性是临时的，因为对数据的业务理解和对问题的理解都会变化的。

## 4. 数据挖掘增大对业务的认知。

数据挖掘是如何产生洞察力的？这有点接近了数据挖掘的核心：为什么数据挖掘必须是一个业务过程而不是一个技术过程？

首先，我们理解业务问题是由人而非算法解决的。数据挖掘者和业务专家从问题中找到解决方案，即从问题的定义域上达到业务目标需要的模式。数据挖掘完全或部分有助于这个认知过程。数据挖掘算法揭示的模式通常不是人类以正常的方式所能认识到的。综合这些算法和人类正常的感知的数据挖掘过程在本质上是敏捷的。在数据挖掘过程中，问题解决者解释数据挖掘算法产生的结果，并统一到业务理解上，因此这是一个业务过程。

其次，这类似于“智能放大器”的概念，在早期的人工智能的领域，AI的第一个实际成果不是智能机器，而是被称为“智能放大器”的工具，它能够协助人类使用者提高获取有效信息的能力。数据挖掘提供一个类似的“智能放大器”，帮助业务专家解决他们不能单独完成的业务问题。

总之，数据挖掘提供一种超越人类以正常方式探索模式的能力，数据挖掘过程允许数据挖掘者和业务专家将这种能力融合在他们的各自的问题的中和业务过程中。数据挖掘过程是一个业务过程。

## 5. 预测模型预测分析提高了信息泛化能力

“预测”已经成为数据挖掘模型（预测模型）可以做什么的可接受的描述，即我们常说的“预测模型”和“预测分析”。这是因为许多流行的数据挖掘模型经常使用“预测最可能的结果”（或者解释可能的结果如何有可能）。这种方法是分类模型的典型应用，比如在我们的DataStudio中，分类算法首先产生含有训练数据的预测模型，然后将预测数据与训练模型相结合进行预测分析，分析预测数据的分类结果。

但是，其他类型的数据挖掘模型，比如聚类和关联模型也有“预测”的特征。这是一个含义比较模糊的术语。一个聚类模型被描述为“预测”一个个体属于哪个群体，一个关联模型可能被描述为基于已知基本属性“预测”一个或更多属性。

同样我们也可以分析“预测”这个术语在不同的主题中的应用：一个分类模型可能被说成可以预测客户行为----更加确切的说它可以预测以某种确定行为的目标客户，即使不是所有的目标个体的行为都符合“预测”的结果。一个诈骗检测模型可能被说成可以预测个别交易是否具有高风险性，即使不是所有的预测的交易都有欺诈行为。

“预测”这个术语广泛的使用导致了所谓的“预测分析”被作为数据挖掘的总称，并且在业务解决方案中得到了广泛的应用。但是我们应该意识到这不是日常所说的“预测”，我们不能期望预测一个特殊个体的行为或者一个特别的欺诈调查结果。

那么，在这个意义下的“预测”是什么？分类、回归、聚类和 关联算法以及他们集成模型有什么共性呢？答案在于“评分”，这是预测模型应用到一个新样例的方式。模型产生一个预估值或评分，这是这个样例的新信息的一部 分；在概括和归纳的基础上，这个样例的可利用信息得到了提高，模式被算法发现和模型具体化。值得注意的是这个新信息不是在“给定”意义上的“数据”，仅有统计学意义。

## 6.数据挖掘的结果的价值不只取决于模型的稳定性或预测的准确性。

准确性和稳定性是预测模型常用的两个度量。准确性是指正确的预测结果所占的比例；稳定性是指当创建模型的数据改变时，用于同一口径的预测数据，其预测结果变 化有多大（或多小）。鉴于数据挖掘中预测概念的核心角色，一个预测模型的准确性和稳定性常被认为决定了其结果的价值的大小，实际上并非如此。

体现预测模型价值的有两种方式：一种是用模型的预测结果来改善或影响行为，另一种是模型能够传递导致改变策略的见解（或新知识）。

 对于后者，传递出的任何新知识的价值和准确性的联系并不那么紧密；一些模型的预测能力可能有必要使我们相信发现的模式是真实的。然而，一个难以理解的复杂的或者完全不透明的模型的预测结果具有高准确性，但传递的知识也不是那么有见地；然而，一个简单的低准确度的模型可能传递出更有用的见解，通俗一点来讲就是：有时，预测结果低准确性的预测模型可能会比预测结果高准确性的预测模型发现更真实的模式，对数据挖掘用户的价值可能也会更大。

 准确性和价值之间的分离在改善行为的情况下并不明显，然而一个突出问题是“预测模型是为了正确的事，还是为了正确的原因？

换句话说，一个模型的价值和它的预测准确度一样，都源自它的业务问题。例如，客户流失模型可能需要高的预测准确度，否则对于业务上的指导不会那么有效。相反的是一个准确度高的客户流失模型可能提供有效的指导，保留住老客户，但也仅仅是最少利润客户群体的一部分。如果不适合业务问题，高准确度并不能提高模型的价值。

模型稳定性同样如此，虽然稳定性是预测模型的有趣的度量，稳定性不能代替模型提供业务理解的能力或解决业务问题，其它技术手段也是如此。

总之，预测模型的价值不仅仅是由技术指标决定的。数据挖掘者应该在模型不损害业务理解和适应业务问题的情况下关注预测准确度、模型稳定性以及其它的技术度量。

## 7.所有的模式（业务目标、数据挖掘目标、预测模型）因业务变化而变化。

数据挖掘发现的模式不是永远不变的。数据挖掘的许多应用是众所周知的，但是这个性质的普遍性没有得到广泛的重视。

数据挖掘在市场营销和CRM方面的应用很容易理解，客户行为模式随着时间的变化而变化。行为的变化、市场的变化、竞争的变化以及整个经济形势的变化，预测模型会因这些变化而过时，当他们不能准确预测时，应当定期更新。

数据挖掘在欺诈模型和风险模型的应用中同样如此，随着环境的变化欺诈行为也在变化，因为罪犯要改变行为以保持领先于反欺诈。欺诈检测的应用必须设计为就像处理旧的、熟悉的欺诈行为一样能够处理新的、未知类型的欺诈行为。

某些种类的数据挖掘可能被认为发现的模式不会随时间而变化，比如数据挖掘在科学上的应用，我们有没有发现不变的普遍的规律？也许令人惊奇的是，答案是即使是这些模式也期望得到改变。理由是这些模式并不是简单的存在于这个世界上的规则，而是数据的反应，这些规则可能在某些领域确实是静态的。

然而，数据挖掘发现的模式是认知过程的一部分，是数据挖掘在数据描述的世界与观测者或业务专家的认知之间建立的一个动态过程。因为我们的认知在持续发展和增 长，所以我们也期望模式也会变化。明天的数据表面上看起来相似，但是它可能已经集合了不同的模式、（可能巧妙地）不同的目的、不同的语义；分析过程因受业 务知识驱动，所以会随着业务知识的变化而变化。基于这些原因，模式会有所不同。

总之，所有的模式都会变化，因为他们不仅反映了一个变化的世界，也反映了我们变化的认知

## 总结

以上所整理的是数据挖掘过程的一些认知，可能大部分已被大家熟知，但肯定仍有一些不甚了解的地方，而最近数据挖掘领域提出的一些新观点也大都和上述各条定律（经验）有关。

最后想告诉各位读者：技术的发展不会改变数据挖掘过程的本质。以上定律（经验）以及这些思想的进一步发展，除了有对数据挖掘者的教育价值之外，还应该被用来判别未来任何数据挖掘过程革命性变化的诉求。