分
计算机世界/２００１年/０５月/２１日/第Ｂ０６版/
专题报道

数据挖掘走入语音处理

王玮　　蔡莲红

数据挖掘技术

数据挖掘(ｄａｔａｍｉｎｉｎｇ)技术是对数据库采取半自动的方式，寻找特定的模式、关联规则、变化规律、异常信息等具有统计意义的结构和事件。自２０世纪９０年代以来，数据挖掘就成为最具活力的研究领域之一，吸引了众多研究者从事这方面的研究。

１．数据挖掘方法的特点

数据挖掘方法与统计方法的不同之处主要体现在：通常的统计方法是在已有的假设基础上，从大量的数据中得到验证，而数据挖掘则是从大量的数据中得到崭新的模式、结论和假设；数据挖掘方法是纯粹的给予数据驱动的方式，而统计方法则更多地引入人为因素并加以分析。探索式数据分析是统计方法中与数据挖掘最相似的分支，但它所面向的数据集还是比数据挖掘对象小得多。

２．数据挖掘的过程

数据挖掘过程可粗略地分为：问题定义(ｔａｓｋｄｅｆｉｎｉｔｉｏｎ)、数据准备和预处理(ｄａｔａｐｒｅｐａｒａｔｉｏｎａｎｄｐｒｅｐｒｏｃｅｓｓｉｎｇ)、数据挖掘(ｄａｔａｍｉｎｉｎｇ)以及结果的解释和评估(ｉｎｔｅｒｐｒｅｔａｔｉｏｎａｎｄｅｖａｌｕａｔｉｏｎ)等阶段。

问题定义：在该过程中，数据挖掘人员必须与领域专家及最终用户紧密协作，一方面明确实际工作对数据挖掘的要求，另一方面通过对各种学习算法的对比进而确定可用的学习算法。后续的学习算法选择和数据集准备都是在此基础上进行的。

数据挖掘：该阶段首先根据对问题的定义明确挖掘的任务或目的，如分类、聚类、关联规则发现或序列模式发现等。之后要决定使用什么样的算法。选择实现算法要考虑两个因素一是不同的数据有不同的特点，因此，需要用与之相关的算法来挖掘；二是要根据用户或实际运行系统的要求，有的用户可能希望获取描述型的(ｄｅｓｃｒｉｐｔｉｖｅ)、容易理解的知识(采用规则表示的挖掘方法显然要好于神经网络之类的方法)，而有的用户只是希望获取预测准确度尽可能高的预测型(ｐｒｅｄｉｃｔｉｖｅ)知识，并不在意获取的知识是否易于理解。

结果解释和评估：数据挖掘阶段发现的模式，经过评估，可能存在冗余或无关的模式，这时需要将其剔除；模式也有可能不满足用户要求，这时则需要整个发现过程回退到前续阶段，如重新选取数据、采用新的数据变换方法、设定新的参数值，甚至换一种算法等。

３．挖掘结果质量评价

类号：ＢＪ８２

数据挖掘结果质量的好坏有两个影响因素：一是所采用的数据挖掘技术的有效性，二是用于挖掘的数据的质量和数据量。如果选择了错误的数据或不适当的属性，或对数据进行了不适当的转换，则挖掘结果不会好。整个挖掘过程是一个不断反馈的过程。

4．主要模型

（1）关联规则模型

发现数据库中数据项之间的相互关系是十分重要的，根据这种关系可以使用户从数据中找到有意义的模式和趋势。以超级市场为例，每个记录包含了一次采购商品的列表，其中关联关系告诉我们两件或更多商品之间的关系。如80%的顾客购买了面包和牛奶，其中有60%的顾客在买面包的同时买了牛奶。我们把这种面包和牛奶之间的关联关系用下列规则方式表示为：面包→牛奶|（60%，80%）。数据项的关联关系也可以在多个项之间产生，例如面包、牛奶→甜酱|（60%，40%）等。目前采用的典型关联算法有Apriori算法和PHP散列表算法等。

（2）神经网络模型

神经网络方法是模拟人脑神经元结构，以MP模型和Hebb学习规则为基础而建立的，主要有三大类多种神经网络模型。

前馈式网络：以感知机、反向传播模型、函数型网络为代表，可用于预测、模式识别等方面。

反馈式网络：以Hopfield的离散模型和连续模型为代表，分别用于联想记忆和优化计算。

自组织网络：以ART模型、Kohonen模型为代表，用于聚类。

神经网络的知识体现在网络连接的权值上，是一个分布式矩阵结构。神经网络的学习体现在神经网络权值的逐步计算上（包括反复迭代或累加计算）。

（3）粗糙集理论模型

粗糙集理论是一种研究不精确、不确定性知识的数学工具，由波兰科学家Z．Pawlak于1982年首先提出。粗糙集的研究主要基于分类。分类和概念（concept）同义，一种类别对应于一个概念（类别一般表示为外延即集合，而概念常以内涵的形式表示如规则描述）。知识由概念组成，如果某知识中含有不精确概念，则该知识不精确。粗糙集对不精确概念的描述方法是：通过上近似概念和下近似概念这两个精确概念来表示。一个概念（或集合）的下近似（lowerapproximation）概念（或集合）指的是，其下近似中的元素肯定属于该概念；一个概念（或集合）的上近似（upperapproximation）概念（或集合）指的是，其上近似中的元素可能属于该概念。粗糙集方法有几个优点：不需要预先知道额外信息，如统计中要求的先验概率和模糊集中要求的隶属度；算法简单、易于操作。

在语音信号处理中的应用

目前，数据挖掘研究主要集中在对新的算法及新的类型的研究上。由于对数据挖掘方法的研究不仅涉及数据挖掘的算法，同时对于需要处理的数据类型也有很高的要求，传统的数

据挖掘的对象主要是超级市场中货篮型数据及经济型数据，几乎很少涉及语音数据的挖掘研究。这一方面是由于语音数据非常复杂，包含很多信息，如基频信息、时长信息、幅度信息、位置信息以及重音信息等，简单来说就是同一个音节在不同的语句中会表现出不同的信息特征，即不同的语境会使音节自身的属性值发生变化，且语音数据是一种时序数据，在一句话中音节的排列是有先后顺序的，同时语音音节之间也存在着很强的音联关系。所有这些信息特征对整个合成系统输出的可懂度以及自然度会产生很大影响。

另一方面，语音数据挖掘的研究需要研究者在语音合成工作积累的基础上才能有效地进行。由于数据挖掘技术对处理对象的要求很高，因此，直接录制音节的波形文件是无法处理的，必须经过严格的预处理过程，如对录音波形进行音节切分和音节标注，这需要大量的人力和物力资源。没有强大的语音处理能力的积累是不可能的。清华大学语音处理实验室长期从事语音信号的研究，具有丰富的语音数据源，即我们通常所说的“熟语料”，这使基于数据驱动的挖掘研究成为可能。将数据挖掘技术应用于语音信号处理可以解决部分现阶段较难解决的语音技术难题，同时尽可能减少人为经验因素对语音处理的影响，完成对语音处理从定性到定量的转变。因此，将数据挖掘方法应用于语音合成具有重要的意义和广阔的前景。

1．关联规则模型获得汉语韵律参数之间的关联关系

语音合成经历了长期的研究发展过程，完成了从实验室向市场应用的过渡，但是，合成系统输出的语音机器味仍然比较浓，与人类自然流畅的发音相比还有较大的差距。这其中主要是受到系统中韵律模块研究的制约，由于韵律模块无法对复杂的韵律特征进行有效描述，因此，合成系统的输出就受到了很大的影响。

韵律特征主要是指音节的时长、基频的包络变化、能量的变化及适当的停顿等众多参数属性，在这些属性中，对合成系统的自然度影响最显著的是音节的基频变化和音长的变化。目前，合成系统中的基频变化规律大多是根据语言学的研究得出的一些定性的描述，这些定性规则能够为合成过程提供一些参考，但是无法在合成过程中直接使用这些规则，而且这些规则也很难覆盖所有的基频变化现象，同时对这些规则的维护和完善也很困难，在具体应用中仍存在较大的不足。由于韵律规则在语音合成中发挥着重要作用，迫切需要采用新的处理方法加以解决。

数据挖掘技术中关联规则模型可以很好地发现数据项之间存在的相互关系，同时有大量的挖掘算法可供选择，因此，基于关联规则的模型可以从大规模语音库中提取更为全面和准确的语音韵律相互关系。首先通过对“熟语料”库中基频数据和时长数据进行预处理，离散化成相应的属性值，获得前后音节的基频信息和时长信息之间的关联关系，从而加以指导合成系统的选音，满足在不同语境下音节参数变化的需求。

2．数据挖掘技术获得汉语韵律的变化规律

在传统的语音研究中，往往是用手工得到语音的基频，求出其调值，然后根据不同情况下调值的变化得到连续变调规律，再将其应用于语音合成系统中进行韵律控制。这是在定性基础上进行的研究，存在很多不足之处。一方面，由于语音数据的变化随机性很大，对少量的语音数据进行处理不能得到较为全面的变调规律，而大量语音数据如果完全用人工来处理

，工作量会很大；另一方面，用人工进行语音数据处理，往往会由于一些先入为主的概念而很难得到较为完全的规律。

基于语音合成中的基音同步叠加技术，可利用数据挖掘技术进行韵律变化规律的学习，采用数据挖掘技术中的神经网络方法、数据项聚类以及粗糙集理论的有机结合进行综合评判，利用神经网络具有的自组织和自学习特性，将经过聚类处理的语音基频数据和时长数据分别转化成神经网络的输入和输出节点，经过网络学习来获得一些典型的基频曲线和时长映射关系。由于神经网络自身理论还存在不够完善的地方，因此，可以辅助以粗糙集理论进行适当的修正，以获得期望的模式。在这些映射的基础上，可通过简单的变换获得典型模式，利用这些典型模式，就可在定量的基础上，对基频的变化规律从较高层次进行韵律规则的研究。

3．基于数据驱动方式的重音确定

在连续语流中，各音节的响亮程度并不完全相同，有的音节听起来比其他音节重，简单地说，这就是重音。以词为考查对象，音位学可划分为正常重音、对比重音和弱重音。人们在口语交流中，常把在表情传意方面较重要的词读得重些，把其余的词读得轻些。语句重音是指由于句子语法结构、逻辑语义或心理情感表达的需要而产生的句子中的重读音，它不同于词重音，因为词重音只出现在词结构中。语句重音一般分为三种：语音重音、逻辑重音、心理重音。

通常研究者认为，重音的声学征兆主要表现在时长、音高与音强三个方面，也往往是三者的结合。不同语言的重音特点不一样，对于汉语，老一辈语音学家赵元任先生认为，“汉语重音首先是延长持续时间和扩大调域，其次才是增加强度。”现代语音学家也认为，汉语重音主要表现在时长的增加（或者说是基音周期数的增加）其次是调域的扩大和音高的提升，调型完整地展开；与发音强度的关系并不是主要的。

以上都是定性的分析，从定性到定量的转换是采用基于数据驱动的方式进行，从大量语料数据本身的特点来分析重音，并且依据重音的特点辅助以韵律学规律，合成更自然的语音信号。

数据挖掘是一种在大量数据库中发现隐藏新知识的计算技术方法。数据挖掘提取的是定性的模型，并且很容易被转化为逻辑规则或用可视化的形式表达。因此，将数据挖掘与人机交互接口紧密联系在一起将对计算机语音信号处理的研究工作产生巨大的推动力，为语音信号处理提供了一条崭新的研究途径。可以预见，采用数据挖掘方法可以较好地解决目前语音信号处理中部分难点问题，从而进一步提高语音合成和语音识别技术的实用化程度。