文章编号：2095-2163(2021)09-0174-04 中图分类号：TP392 文献标志码：A

智慧数据库系统上的多领域特征表征与综合

张翔熙，王宏志
（哈尔滨工业大学 海量数据计算研究中心，哈尔滨 150001）

摘 要：在智慧数据库系统预测任务中，查询的向量表示往往需要多个领域的特征共同作用，才能得到较好的效果。针对多领域特征的表征与综合问题，本文通过深度神经网络的理论与技术，为语义、结构等领域特征设计了张量化的表示方案，并提出权基综合、感知机综合、虚线综合 3 种多领域特征综合方案。大量真实数据的实验结果表明，所提出的多领域特征与综合方法，能够有效地提取与转化多领域的查询相关特征，具有较好的收敛效果，能够为智慧数据库系统的其它预测任务提供查询向量表示方面的支持与嵌入。

关键词：智慧数据库；深度神经网络；多模态；特征综合

Multi-field feature synthesis in smart database system

ZHANG Xiangxi, WANG Hongzhi
(Massive Data Computing Research Center, Harbin Institute of Technology, Harbin 150001, China)

【Abstract】 In the prediction task of a smart database system, the vector representation of a query often requires a combination of features from multiple fields to achieve better results. Aiming at the problem of representation and synthesis of multi-domain features, this paper uses the theory and technology of deep neural networks to design a quantitative representation scheme for domain features such as semantics and structure, and proposes three kinds of multi-domain synthesis, i.e. weight-based synthesis, perceptron synthesis, and jump line synthesis. Experimental results on a large amount of real data show that the proposed multi-domain feature representation and synthesis method can effectively extract and transform multi-domain query-related features which has a good convergence effect and can provide query vectors for other prediction tasks of smart database systems.

【Key words】 smart database system; deep neural network; multimodal; feature synthesis

0 引 言

作为现代软件系统中至关重要的一部分，数据库系统一直在软件系统中为数据存储、数据控制与数据分析提供关键的支撑。然而，随着数据库系统工业实践日趋复杂、理论基础日趋完备，现代数据库系统往往有成百上千个可选配置参数与调优选项，对数据库管理员的心智负担也日趋沉重；同时，大规模分布式高吞吐量的现代数据库系统应用，也对数据库查询存储优化等问题提出了更高的要求。

为了解决这些问题，以深度学习为代表的统计学习方法的智慧数据库系统技术应运而生。借助统计学习技术，智慧数据库系统在配置、优化、设计、监测等多个子领域上分别开辟了新的研究方向与研究热点。智慧配置能够显著降低数据库管理人员心智负担，降低因配置不当造成的资源开销与浪费；智慧优化能够在过去数据库系统优化理论的基础上进一步突破，通过统计学习的手段来解决传统方法因复杂度、近似比等理论限制而难以完成的优化任务；智慧设计能够根据不同的工作流，自适应地改变索引与存储的数据结构；智慧监测通过时间序列分析等手段，自动提前发现运行异常，规避运维风险。

在智慧数据库系统中的一个典型任务，就是对于来自语义、数据库结构、运行环境等多个领域的特征进行向量化的表征与综合。通过通用化的建模手段，解决智慧数据库系统上多领域特征与综合中的问题与困难，将能够作为底层支撑技术，为更加复杂的数据库系统模型提供必要的特征侧支撑。

本文针对智慧数据库系统上的多领域特征表征与多领域特征综合两个关键问题，建立了一套能够综合结构、语义等特征的深度神经网络结构与体系，并在具体的预测任务与百万级真实样本上进行了验证与评估。该系统具有一定的可扩展性，可以作为底层结构参与到更加复杂与困难的智慧数据库系统

基金项目：CCF-华为数据库创新研究计划项目(DBIR2019005B)。

作者简介：张翔熙(1996-)，男，硕士研究生，主要研究方向：智慧数据库系统。
通讯作者：王宏志 Email：wangzh@hit.edu.cn

收稿日期：2021-06-09

模型之中。

1　多领域特征的表征

多领域特征表征(multi-field feature representation)是智慧数据库系统中的常见问题。与图像识别等传统单领域的任务不同,数据库系统中某个查询的执行情况与查询语句、数据结构、运行环境等多个领域的信息都有关联。为了能够通过深度学习等统计学习手段,解决智慧数据库系统上的预测任务,这些多领域的特征必须通过适当的处理、解析成能够被神经网络模型处理与分析的张量形式。对此,本文将多领域特征分为语义特征、结构特征、辅助特征 3 类,分别进行向量化表征工作。

语义特征是指以 SQL 语句为代表的查询任务描述中包含的任务信息。对于语义特征,首先过滤掉 SQL 语句中的数字及尾部分号,保留包含表名、属性名、保留字、运算符在内的 134 个词构成的词表。然后,对于单条 SQL 语言输入,按照输入中各个词的位置,得到一个 multi-hot 编码的、各个位置代表词出现次数的 134 维向量。将该向量经过一个可训练的词向量矩阵变换后,得到 32 维的、包含了查询语义特征的张量输出(在综合时称为 ***lexical_embedding***)。

结构特征是指查询所指涉的数据表、属性集、选择条件包含的任务信息。对于结构特征,先把查询所在表的 ID 进行编码与嵌入,得到 61 维(与属性总数相同)的 table_embedding 张量结果;再使用直方图预估的方式,对于查询的 WHERE 子句中,由 AND 连接的每一个属性选择条件,独立地估计该条件所筛选的元组比例(selectivity),组成一个与属性总数目等长的 61 维向量 range_vector,将其与 table_embedding 逐位相乘后,得到 ***cross_embedding*** 输出;另外,通过常量的方式计算一个 61 维的 ***mask*** 向量,使得其中只有该表对应的属性位置填写该表的总行数,再与前述的 ***range_vector*** 逐位相乘,获取 ***hist_embedding*** 输出。将 3 个输出合并,就得到了代表结构领域特征的 183 维的张量输出(在综合时称为 ***structure_embedding***)。

除结构与语义方面的特征之外,在系统运行的过程中,往往有其它的信息,同样影响预测任务的输出结果,这些结果可以通过连加或连乘的方式进行综合,得到可以辅助训练的稠密特征输入。在实验任务中,通过常量的手段按照查询涉及的表来获取对应表的静态总数,与前述的 ***range_vector*** 进行折叠相乘,就得到了一个一维的标量输出(在综合时称为 ***cross_bias***)。

经过上述的操作,就将多种不同领域的特征,通过网络操作进行综合,分别得到了 3 个不同领域的输出张量。本文设计的方式能够解决多领域特征的表征问题,并为多领域特征的综合问题提供了技术基础;相应方法的有效性将在实验章节中得到进一步的验证。

2　多领域特征的综合

在底层获取了 3 个不同途径的多领域特征表征之后,还需要解决多领域特征的综合问题。要想实现多种不同领域特征的综合,简单直接相连或加和可能会带来预测性能上的严重损失,往往需要设计符合问题性质的综合结构。

为了验证不同综合结构对于预测性能带来的影响,本文设计并实现了 3 种彼此不同的多领域特征综合方式,并通过实验手段,探究不同综合方式对于智慧数据库系统上预测问题带来的影响。

权基综合是相对最直观的多领域特征综合方式。由于各个领域对于最终预测目标的贡献不同,因此使用 3 个可学习的权重变量,通过 softmax[1] 转化为 3 个总和为 1 的非负权重后,来综合各个领域的预测结果。其中,单个领域的预测结果,对于结构领域和语义领域,通过 ELU[2] 激活的单层感知机网络进行单目标的预测;对于辅助稠密特征,则使用可训练的线性变换进行预测。这样,3 个预测结果在 3 个非负权重的加权下,得到最终用于计算损失的输出。

感知机综合是通过神经网络的结构,对输入的 3 个领域特征进行进一步的抽象与交叉。对于上节中输出的 ***lexical_embedding***、***structure_embedding*** 与 ***cross_bias*** 3 个张量,通过串接的方式得到一个长张量,作为感知机的输入。随后,通过 ELU 激活的多层感知机进行处理与输出,最终得到单目标的预测结果。这种方式在综合时引入了更多的非线性因素,提高了网络的综合能力。

虚线综合则从损失函数的角度,考虑多领域特征的综合问题。首先,使用与权基综合相似的方式,让每一个领域都给出一个单目标的预测输出;再通过 3 个可学习的权重,将预测输出结果进行线性加权。然而,简单的线性加权会导致部分领域预测一个负数,而其它领域输出巨大正数的现象,这种拮抗会降低系统的稳定性。因此,对于每一个领域的预

测结果在加权前,通过虚线连接到样本标签,计算一个用于辅助的均方误差;在训练时,将辅助误差与真实误差进行加权再梯度下降。这样,通过虚线误差能够强制每个领域进行稳定的训练,提高预测的精准性与稳定性。

3 实验

为了验证本文提出的多领域特征表征与综合方法的有效性,对比 3 种多领域特征综合方法的优劣,本节将在真实运行环境下收集样本并进行训练比对。

实验样本的生成,依赖的数据表为 tpch-gen[3] 标准开源程序所生成的、存储在 MariaDB[4] 开源数据库系统上的 tpc-h[5] 标准数据库。包含 8 个表、61 个属性、356 万行记录,其主键、外键、索引等都符合 tpc-h 的规范。在此基础上,通过程序随机生成各个表上的 SQL 查询语句。其 FROM 部分随机为 8 个表中的任意一个,SELECT 部分任意随机选择该表上的任意数量属性,WHERE 部分随机生成随机数目表上合法属性的等值或不等值查询子句构成合取式。通过这样的操作,生成了 77 万条查询语句,并在单机的 MariaDB 上分别运行并计时。

经过上述过程,收集到 77 万条样本,将模型在 TensorFlow[6] 框架实现后,基于纯 CPU 的运行环境分别进行训练与误差统计。对于只使用语义信息(lexical_only)、结构信息(structural_only)、权基综合(softmax_model)、感知机综合(deep_cross)、虚线综合(boost_model)等 5 个不同模型的运行结果,分别按照迷你批次序号,与对应的损失函数值绘制图像,如图 1 所示。

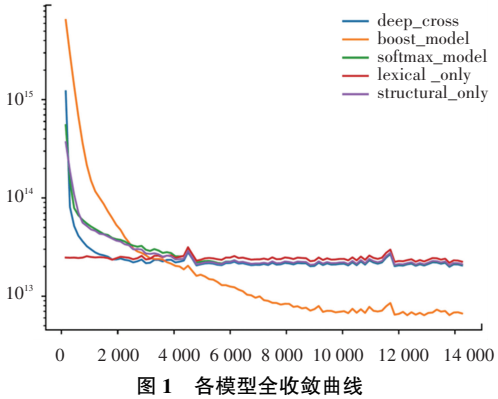

图 1 各模型全收敛曲线

Fig. 1 Convergence curves of all models

由于部分图像不够清晰,对于图 1 中几个在 2×1 013位置收敛的曲线、放大尾部部分显示如图 2 所示;对于前4 000个迷你批次,放大头部部分显示如图 3 所示,以观察不同模型收敛效率的区别。

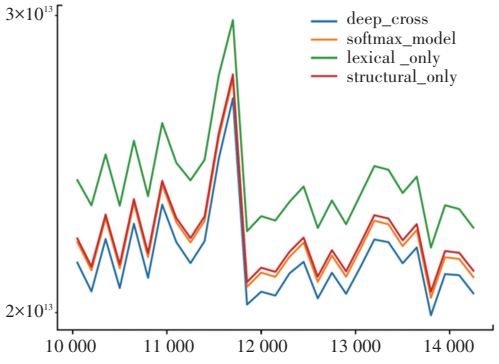

图 2 收敛曲线尾部放大图

Fig. 2 Tail part illustration of convergence curves

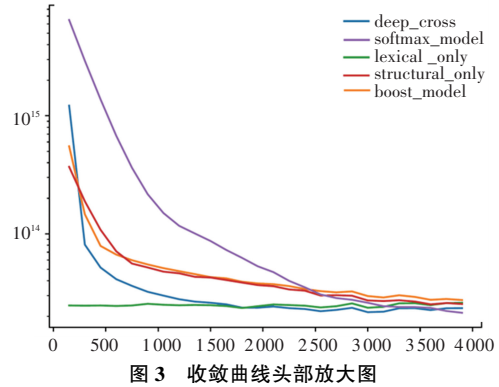

图 3 收敛曲线头部放大图

Fig. 3 Head part illustration of convergence curves

分析以上数据,可以得到以下结论:

(1)各曲线都能够有效收敛。可见本文提出的多领域特征表示方式,能够从输入特征中提取出有效信息,让查询时间预测这一常见的智慧数据库问题有较好的实验结果,验证了第一节内容的有效性;

(2)从图 2 中可以清晰地看到:不论是语义信息还是结构信息,如果只使用单独一个领域的信息,在最终的模型效果上都会明显弱于多领域模型,证明了多领域特征综合的必要性;

(3)3 个多领域综合方式都能够有效地进行特征综合,在最终效果上虚线综合最好,感知机综合其后,最后是权基综合,证明了本文提出的方法的有效性和创新性;

(4)在收敛速度上,感知机综合模型收敛非常快,最终收敛位置第二优;虚线综合虽然明显最终效果更好,但是由于辅助损失对于主要损失在前期有干扰作用,收敛速度最慢。因此,在小样本条件下,感知机综合效果更好;大样本量下,则虚线综合方式更优。可见,本文提出的两种多领域特征综合方式之间,能够根据现实场景互补,具有较好的适应性。

(下转第 183 页)